The
University
Of
Sheffield.

# Ethical decision system for autonomous vehicles in unavoidable accident scenarios

**Lalis Millan Blanquel**

A thesis submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy

Automatic Control and Systems Engineering
The University of Sheffield
United Kingdom

December 2021

*To my parents.*

# Acknowledgements

First, I want to thank my supervisor Prof Robin Purshouse for the guidance and support that he provided me during the realisation of this thesis. Also, to Prof Sandor Veres for his help and valuable suggestions.

I would like to express my sincere gratitude to my mother for always believing in me, thanks for your endless love and support. My deepest gratitude goes also to my father, I would not be the person that I am today if it was not for your care, inspiration, and advice.

Words cannot express my gratitude to my dear husband Fernando. Thank you for being my best friend, partner, and advisor. All of this is only possible because of your support, guidance, and care. You gave me the encouragement I needed to achieve this endeavour.

Thanks should also go to my friends back in Mexico and all the new friends that I made in Sheffield. I appreciate your friendship and reassurance in this path I decided to take. Also thanks to all the ACSE staff that allowed me to work as a GTA during my time here, more specifically Tara Baldacchino and Roderich Gross for the opportunity to work with them. I would like to acknowledge Rafael Cardoso from the University of Manchester for helping me understand Jason and my special thanks to Phebe Bonilla for her invaluable help proofreading this work.

Lastly, I would like to mention the Mexican National Council of Science and Technology (Consejo Nacional de Ciencia y Tecnologia) for the financial support that made my postgraduate studies possible (CONACyT Scholarship No. 440758).

# Declaration

I hereby declare that the present thesis was composed by myself in its entirety and the work contained herein is my own except when stated otherwise. Chapter 3 was done in collaboration with Prof Robin Purshouse and Prof Sandor Veres. Chapter 3 was also presented at the *2020 Mediterranean Conference on Control and Automation*

# Abstract

The automotive industry is heading towards the introduction of fully autonomous vehicles. However, before this type of vehicles are commercially available at mass scale, some issues need to be solved. A major issue is the ethics involved in the decision-making during an accident; this research presents an analysis of how to approach this issue and a way to implement a solution based on concepts from *Belief-Desire-Intention* agent modelling. The first part of this research identifies and defines a pre-programmed system with different ethical settings based on five formal ethical theories. For each, eight ethical concerns are defined and ordered accordingly. These concerns are defined in terms of harm to self and harm to others. The ethical concerns are used as a guideline to define the level of importance of each person or object in an accident scenario. The resulting rank of concerns is novel in the field of ethical decision-making for autonomous vehicles and serves as basis for the implementation of an ethical decision-making agent for unavoidable collisions. The second part of this thesis focuses on the design, derivation and implementation of the decision-making system in a BDI agent framework. With the proposed system, the vehicle is partially tailored to the ethical preferences of different users while still being bounded by legal requirements to avoid any misuse. The resulting outcomes of the decision-making system under different scenarios are shown and discussed. In these discussions it is clear that the proposed system successfully captures ethical concerns, priorities and behaves in accordance to the ethical theories.

# Contents

**1 Introduction** **1**

1.1 Motivation . . . 1

1.2 Overview and Main Contributions . . . 2

1.3 Associated Publications . . . 3

**2 Literature Review** **4**

2.1 Ethical Considerations . . . 4

2.1.1 Ethical Theories . . . 4

2.1.2 Ethics and Autonomous Vehicles . . . 5

2.1.3 Ethical perspectives in the context of AV . . . 7

2.1.4 Surveys . . . 7

2.1.5 Proposed Implementations . . . 9

2.2 Legal Considerations . . . 10

2.3 Advanced Driver-assistance Systems (ADAS) . . . 12

2.3.1 Warning systems . . . 13

2.3.2 Automated lane change . . . 14

2.3.3 Vehicle navigation . . . 15

2.3.4 Naturalistic driving . . . 16

2.3.5 Fully autonomous vehicles . . . 17

2.3.6 Collision avoidance . . . 18

2.3.7 Existing Proposals . . . 19

2.3.8 Conclusion . . . 20

**3 Ethical Perspectives** **23**

3.1 Ethical decision-making System . . . 23

3.2 Ethical Concerns . . . . . . . . . . . . . . . . . . . . . . . . . 24
3.2.1 Discussion . . . . . . . . . . . . . . . . . . . . . . . . 26
3.3 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . 27

**4 Decision-making system 28**

4.1 Proposed System . . . . . . . . . . . . . . . . . . . . . . . . 28
4.2 Agent Architecture . . . . . . . . . . . . . . . . . . . . . . . 31
4.3 BDI agents . . . . . . . . . . . . . . . . . . . . . . . . . . . 31
4.4 AgentSpeak and Jason . . . . . . . . . . . . . . . . . . . . . . 33
4.5 Ethical Decision-Making Agent . . . . . . . . . . . . . . . . . 35
4.5.1 Environment perception . . . . . . . . . . . . . . . . . . 36
4.5.2 Decision-making . . . . . . . . . . . . . . . . . . . . . . 38
4.5.3 Manner of Collision . . . . . . . . . . . . . . . . . . . . 40
4.5.4 Crash Severity . . . . . . . . . . . . . . . . . . . . . . . 42
4.5.5 Etical Decision-making framework . . . . . . . . . . . . . 42
4.5.6 Utilitarian decision-making . . . . . . . . . . . . . . . . 44
4.5.7 Distributive Justice decision-making . . . . . . . . . . . . 45
4.5.8 Kantian decision-making . . . . . . . . . . . . . . . . . . 46
4.5.9 Altruism decision-making . . . . . . . . . . . . . . . . . 47
4.5.10 Ethical Egoism decision-making . . . . . . . . . . . . . . 48
4.6 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . 49

**5 Simulation and scenario analysis 50**

5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . 50
5.2 Vehicle and Environment Implementation . . . . . . . . . . . . 50
5.3 Scenarios Definition . . . . . . . . . . . . . . . . . . . . . . 56
5.4 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59
5.4.1 Scenario I . . . . . . . . . . . . . . . . . . . . . . . . 59
5.4.2 Scenario II . . . . . . . . . . . . . . . . . . . . . . . . 60
5.4.3 Scenarios IIIA and IIIB . . . . . . . . . . . . . . . . . . 66
5.4.4 Alternative Severity Score Calculation . . . . . . . . . . . 80
5.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . 80
5.6 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . 82

**6 Conclusions** **84**

6.1 Conclusions and discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84

6.2 Societal impact . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 85

6.3 Limitations and Future Directions . . . . . . . . . . . . . . . . . . . . . . . . 86

# List of Figures

4.1 Proposed system. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29

4.2 Flow diagram of the proposed system. . . . . . . . . . . . . . . . . . . . . . . 30

4.3 Architecture of the ethical agent and simulation . . . . . . . . . . . . . . . 31

4.4 Jason Reasoning Cycle [1] . . . . . . . . . . . . . . . . . . . . . . . . 34

4.5 Overview of the ethical decision-making system using Prometheus notation. 36

4.6 Class diagram of the ethical decision-making environment. . . . . . . . . . 37

4.7 Internal action class diagram. . . . . . . . . . . . . . . . . . . . . . . . 39

4.8 Angle Collision. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41

4.9 Head on collision. . . . . . . . . . . . . . . . . . . . . . . . . . . . 41

4.10 Rear-end collision. . . . . . . . . . . . . . . . . . . . . . . . . . . . 41

4.11 Sideswipe collision. . . . . . . . . . . . . . . . . . . . . . . . . . . 41

4.12 Ethical agent framework. . . . . . . . . . . . . . . . . . . . . . . . . 43

5.1 System Matlab model . . . . . . . . . . . . . . . . . . . . . . . . . . 51

5.2 Coordinate systems for the simulation. . . . . . . . . . . . . . . . . . . . . 51

5.3 3-DOF Bicycle model used for simulation. . . . . . . . . . . . . . . . . . . 52

5.4 System communication diagram. . . . . . . . . . . . . . . . . . . . . . . . 55

5.5 Scenarios to test the system. The orange car represents the ego vehicle, the red arrows represent the direction in which the actors are moving. . . . . . 57

5.6 Initial physical conditions for all actors . . . . . . . . . . . . . . . . . . 58

5.7 Utilitarian theory result scenario I. Turning left is the option that has the lowest severity score, hence the ego vehicle turns left. The results obtained with all the others ethical settings is the same, always turns to the left. . . . 60

5.8 Utilitarian theory result scenario II. In this scenario the pedestrians (in black and green) can not be avoided without harming the passengers and since they are the priority the vehicle decides not to change the direction, harming only one person. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 61

5.9 Distributive justice theory result scenario II. . . . . . . . . . . . . . . . . . . 62

5.10 Kantian theory result scenario II. In this situation, turning left is the only option where the vehicle can avoid the pedestrians, hence this is the selected option. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 63

5.11 Altruist theory result scenario II. Even though turning to the left is more dangerous for the passengers, the altruist agent decides to protect the pedestrians. 64

5.12 Egoist theory result scenario II. . . . . . . . . . . . . . . . . . . . . . . . 65

5.13 Utilitarian theory result scenario IIIA. The vehicle priority is the pedestrians (green and black), from the two options available the less risky is to turn to the right. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 66

5.14 Utilitarian theory result for scenario IIIB. The vehicle priority is the passengers, reason why it decides not to change of direction, harming the pedestrians instead of the passengers. . . . . . . . . . . . . . . . . . . . . . . . . 67

5.15 Distributive justice theory result scenario IIIA.The manoeuvre where less people is involved and has the smallest severity score is turning to the right. 68

5.16 Distributive justice theory result scenario IIIB. In this scenario configuration, even though there are more people involved in the collision, if it turns right, this option is still the best one as it involves less people than going forward and it severity score is smaller that going left. . . . . . . . . . . . . . . . 69

5.17 Kantian theory result scenario IIIA. . . . . . . . . . . . . . . . . . . . . . 71

5.18 Kantian theory result scenario IIIB. . . . . . . . . . . . . . . . . . . . . . 72

5.19 Altruist theory result scenario IA. . . . . . . . . . . . . . . . . . . . . . . 74

5.20 Altruist theory result scenario IB. . . . . . . . . . . . . . . . . . . . . . . . 75

5.21 Ethical Egoism theory result scenario IIIA. . . . . . . . . . . . . . . . . . 77

5.22 Ethical Egoism result scenario IIIB. . . . . . . . . . . . . . . . . . . . . . 78

# List of Tables

3.1 Ethical concerns ranked according to each ethical theory from highest priority at the top to lowest priority at the bottom . . . . . . . . . . . . . . . . . . 25

4.1 Crash severity . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42

4.2 Priority score values for avoiding objects . . . . . . . . . . . . . . . . . . . . . 44

5.1 Vehicle parameters used for simulation. . . . . . . . . . . . . . . . . . . . . . 54

5.2 Ego vehicle initial conditions for all scenarios. . . . . . . . . . . . . . . . . . 58

5.3 Actors initial conditions for all scenarios. . . . . . . . . . . . . . . . . . . . 58

5.4 Severity scorer results of scenario I for all agents . . . . . . . . . . . . . . . 59

5.5 Final results for all ethical agents in all the different scenarios. . . . . . . . 79

# List of Algorithms

1 Utilitarian decision-making . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45
2 Distributive justice decision-making algorithm . . . . . . . . . . . . . . . . . . 45
3 Kantian decision-making algorithm . . . . . . . . . . . . . . . . . . . . . . . . 46
4 Altruist decision-making . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 47
5 Ethical Egoism decision-making . . . . . . . . . . . . . . . . . . . . . . . . . 48

# Acronyms

**ADAS** Advanced Driver-assistance Systems.

**AVs** Autonomous Vehicles.

**BDI** Belief Desire Intention.

**MIO** Most Important Object.

**NHTSA** National Highway Traffic Safety Administration.

**PAX** Passenger.

**POTV** People Outside The Vehicle.

**TCP** Transmission Control Protocol.

# Chapter 1

# Introduction

In recent years, there has been an increasing interest in the development of Autonomous Vehicles (AVs) . Academy and industry have taken different approaches. Recently, the competition between automakers to commercialise AVs has amplified the interest in addressing different issues stated by governing authorities and the public.

One of the concerns about the introduction of autonomous technology is related to the ethical and legal aspects of the behaviour of the vehicle in an emergency. The main issue is to decide what manoeuvre should the vehicle perform in a situation where a collision cannot be avoided? Further more, who is to decide it? The programmer, the carmaker or the government? And who will be accountable for the the consequences?

In the following sections, the motivation for this work, the objectives and contributions are presented.

## 1.1 Motivation

Despite the increasing interest in AVs, from the literature can be identified that there is little research on how to tackle the moral and legal questions derived from their use. Simultaneously, from the engineering perspective, there is lack of emphasis on this predicament. These issues have been mostly researched from a philosophical, legal and a psychological point of view; however, there is little understanding about how to translate the findings to an engineering solution.

Another gap identified in the literature is that most of the research is focused on how to avoid crashes, but does not provide a solution as to what to do in case of an unavoidable accident. This is particularly important, as it is difficult to guarantee that any autonomous vehicle is going to be 100% secure [2, 3], especially during their first interactions with regular vehicles and people who are unfamiliar with this technology.

Equally important is the lack of emphasis on solutions to the moral questions surrounding autonomous vehicles. Studies related to user preferences show that the utilitarian approach has a good acceptance level, though most subjects also agree that they would not buy a utilitarian car. Furthermore, they are not willing to accept governmental enforcement of this approach [4] [5]. In addition, surveys found in the literature are usually not neutral; their questions tend to be too specific about the type of pedestrian (children, elderly people, different gender or occupation and friends) which can cause unwanted biases [6] [7].

## 1.2 Overview and Main Contributions

The aim of this research is to develop a decision-making system for autonomous vehicles that defines their behaviour in case of an emergency in which a collision cannot be avoided. Parameters to consider consist of passenger preferences, the characteristics of the vehicles as well as their dynamics, the objects and pedestrians in the vicinity and safety systems present in the vehicle. Particular objectives include:

- Perform a literature review of the different ethical approaches, existing legal frameworks and Advanced Driver-assistance Systems (ADAS)
- Identify different ethical perspectives that can be applied to autonomous vehicles.
- Define the algorithms to implement the selected ethical perspectives.
- Design the decision-making system, implement it and test it .
- Evaluate the system in terms of the inputs to the system, the real-time decision process and compliance with current legislation requirements.

This thesis is organised as follows. In Chapter 2 we present the definition of the ethical theories that we are using and some legal considerations. Chapter 3 introduces the ethical perspectives employed in this research. Chapter 4 covers the design of the ethical decision-making agent. In Chapter 5 we present the environment simulation, the testing scenarios and the outcome for each ethical theory. Finally, Chapter 6 provides a summary of the results and a discussion of the limitations and future directions for this work.

## 1.3 Associated Publications

Part of this work has been published in:

### Conference Proceedings

1. L. Millán-Blanquel, S. M Veres, and R. C. Purshouse, **"Ethical considerations for a decision-making system for autonomous vehicles during an inevitable collision"**, *2020 28th Mediterranean Conference on Control and Automation (MED)*, 2020, pp. 514–519.

# Chapter 2

# Literature Review

In this chapter we present the state of the art and previous works that motivate this research. To the best of the author's knowledge, systems like the one proposed in this work were not published at the time of starting this work. However, during the course of this research, some examples have been published. We discuss them at the end of this chapter.

The literature review is separated in three parts: Ethical considerations. legal considerations and Advanced driver-assistance systems. In each of these parts we touch on several concepts and tools that we have used to derive the results of the present research.

## 2.1 Ethical Considerations

Autonomous vehicles pose several challenges including technological developments and ethical dilemmas. These dilemmas must be solved in order to AVs to be implemented and accepted by society. In this section, we discuss the ethical theories that can aid in this task.

### 2.1.1 Ethical Theories

In the philosophy field, there exists a number of interpretations for ethics. Ethics are moral philosophies that encompass rules and behaviour standards that can serve as a guide for decision-making [8]. Different ethical perspectives are encountered through the literature; however not all of them may be applicable to AVs. In this section we present the main five that we have selected for our system. A more comprehensive explanation of these theories can be found in [9, 10, 11].

*Utilitarianism* looks for the greater good. It defines the morality of an act in terms of its consequences; if a decision produces good for the most people, then it is morally correct [10]. Although utilitarianism may be the solution preferred to implement in AVs, one of the problems with it, it is that it does not solve the question of what to do when the number of persons harmed is equal for both decisions. Imagine a situation where the problem is to decide if the AV should save its passenger or a pedestrian. In this situation utilitarianism does not provide a clear answer.

*Deontology* is a moral theory focused on duties and rights, with an emphasis on the action itself and not the result of said action. This contrasts to other theories like utilitarianism, where the focus is on the outcome of the action [12]. Different authors have given their interpretation of deontology; one of the most known is by Kant.

*Distributive justice* is a theory developed by John Rawls where he emphasizes a fair and equal allocation of resources [11, 13]. The key to this theory is to place the self in what he calls 'the original position' this means taking a perspective where the relative merits of each position or role in the scenario are understood, but from these positions, the one occupied by the self is unknown. Since the self is ignorant of its position, Rawls argues that the situation can be analysed without biases with the objective to optimise the worst outcome, as that could be the self's outcome itself.

*Altruism* is to care for the well-being of the others, even if that may lead to self-sacrifice to protect other people [10]. This theory can be applied to AVs. However, not all potential buyers may agree to ride on a vehicle that not prioritise their safety.

*Egoism* can be divided into two types. The first one is psychological egoism, where decisions are made based on a natural impulse of self-satisfaction and instant gratification. The second one, ethical egoism, argues that the self should seek its happiness using its rational judgment, deciding according to principles such as integrity and honesty. As opposed to altruism, some philosophers argue that caring for one's self is also morally acceptable [12, 9, 13]. While egoism is a theory that might be suitable for AVs, it will require legislation, as it may not be acceptable to cause harm to many in order to save one person. For our system, we will implement an ethical egoism approach.

### 2.1.2 Ethics and Autonomous Vehicles

Imagine a situation where five persons are unexpectedly on the road and there is not enough time to stop the vehicle. The only option is to steer the vehicle, but in doing so, it would hit another person. This is known as the trolley problem [14]. This dilemma has been widely used in the literature to approach the problem of ethics in autonomous vehicles. In [7] it is explained the importance of ethical decision-making for AVs. A proposed solution to the dilemma is to give back the control of the vehicle to the driver. However, this may be impossible in some situations due to the time required for the driver to regain control, analyse the situation and react accordingly. Moreover, this solution is not compatible with the goal of having fully automated vehicles.

While driving, humans make decisions based on their knowledge and experience. During an emergency, stress, time constraints, and emotions are factors that influence the decision-making process. However, when a machine makes a decision, it is based on calculations and logical processes. Machines do not feel stress and can analyse a situation faster than a human can. Autonomous vehicles would make these critical decisions without the intervention of a human. However, when faced with a situation that involves human lives, society needs more that a set of probabilities to justify and understand why certain decisions were taken [15].

The conditions to decide who would be the one to be harmed have also been debated. To decide between hitting one person or another, using criteria as age, race, gender, profession or disability can be considered a discrimination issue, and is prohibited not only by governments [16] but also by Engineer associations [17].

Using a utilitarian approach is constantly discussed between scholars. The question with these approaches is to define what is more important, to preserve the liberty and the individual rights of each person, or to seek the greater good for the community. In [18], the author points out that the user of the vehicle should be able to decide the ethical behaviour of the vehicle, as may not be a good or wrong answer in setting moral behaviours. However, some researchers [19] defend a mandatory ethical setting over a personal one, arguing that most people would select an egoistic behaviour that would make the situation worst for everyone. Hence, a mandatory ethical setting would benefit everyone. The guidelines for designing the decision-making algorithms should be carefully reviewed to ensure that no hidden discrimination or parameters that can bias the decision of the AV are present to the detriment of a certain group of people [20].

Although, the trolley problem has been popular in the debate over ethical questions about AVs, some authors argue that this abstraction is inappropriate [21]. One difference between a trolley problem scenario and AVs is that, in the case of the AVs, the decision was made a long time before the accident, whereas in a trolley problem the person is forced to make the decision in real-time. Additionally, in the case of AVs the decision needs to be made with a thorough analysis of the situation and should be agreed upon by various stakeholders (e.g. car manufacturers, system developers, society, lawmakers). Another point of comparison is that the outcome of the trolley problem is assumed to be completely certain, whereas AVs only work with estimates and should consider risk assessment and decision-making under uncertainty to define the best actions.

While it is true that the trolley problem may be a simplistic case, its use in research is useful to develop safety features, as there are suggestions that autonomous vehicles would not be 100% safe [22]. Hence the importance on developing systems that can help to resolve these ethical issues.

### 2.1.3 Ethical perspectives in the context of AV

Some of the mentioned ethical theories have been discussed in the context of AV in the literature. Utilitarianism is the most popular suggestion for implementation in AVs [18, 19]. One of the problems with this scheme is that it does not solve the question of what to do when the number of persons harmed is equal for all possible outcomes. Take, for example, a situation where the problem is to save the passenger of the vehicle or to save a pedestrian. In this situation, utilitarianism does not provide a clear answer.

Altruism is another theory that has been implicitly discussed in relation to AVs [4]; however not all potential buyers may agree to ride in a vehicle that will not prioritise their safety. Egoism may also be suitable, but it requires legislation as it may not be acceptable to cause harm to many people just to save one person. In general, it is not clear that any single ethical perspective can universally satisfy all possible sets of preferences or scenarios.

### 2.1.4 Surveys

To solve the ethical question about AVs, researchers need to know the societal views on these moral questions. Some surveys have been conducted to understand what actions people believe are morally correct. In [4] six surveys that shed light on this issue are presented. Through an online platform they asked the participants what would be the moral course of action in a trolley problem situation, where the vehicle should decide between protecting one pedestrian or protecting several pedestrians. In this scenario, the majority agreed that the vehicle should protect the greater number of lives. However, in a variation of the problem, where the car should decide between the driver or the pedestrians, with just a 23% of approval, they do not agree to sacrifice the passenger to save just one life, although from two to more pedestrians saved, the approval ratings increased. Interestingly, when the participants were asked if they would buy a car programmed to behave in a utilitarian way, the majority said they would not. Further questions relate to governmental enforcement of utilitarian behaviours, again, they agreed that it is morally acceptable to sacrifice the passenger but they do not approve of having this behaviour legally enforced. These results are consistent with those on [5]. On the first scenario of their survey, they present the participants the classic trolley problem, obtaining a 95.4% approval to save more than two lives. However, in the scenario in which the driver is sacrificed to save just two lives, the approval rating decreases to 52%. While the utilitarian behaviour still has a high acceptance approval, it is clear how people's perception changes when it is their own life at risk. Similarly, most participants on [23] preferred the utilitarian decision; however, this study did not ask participants whether they would buy a utilitarian vehicle or not. In their second experiment participants were asked who they think should have the responsibility for taking a decision in a dilemma situation; the majority said that ethics researchers carry a greater obligation to define how AVs should behave.

The already mentioned surveys, study what people think is the correct decision and how a vehicle should behave. However [24] highlights the fact that, people are averse to allowing machines to make moral decisions. This is caused by the perception of a lack of emotions and expertise of the machines. The authors found that, when comparing a human to a machine making the same decision with the same outcome, the participants always deem it more permissible for the human to make the decision. This is regardless of the positive or negative impact of said decision. The only circumstance in which the participants preferred the machine to make the decision is when it is explicit that the machine has a higher rate of success, for example taking medical decisions. If it is not clearly stated in a pair-wise comparison, people prefer an average human to take the decision, even if the machine is more expert. The authors also suggest that people are more open to accepting decisions made by humans aided by machines, but ultimately the human should make the decision. By contrast, [25] reports that human aversion to machines making moral decisions could be intuitive and hard to rationalize. While this study also concludes that participants preferred humans to make decisions, the authors were not able to point out a reason for this aversion. They argue that it is important to continue research within this field and that the industry should not discount people's moral aversion to machines, as this could negatively affect the AVs market. As mentioned in [26], ethical concerns could be a factor of resistance for the acceptance of autonomous vehicles.

Another aspect that must be addressed is cultural differences; car sales are a global market with each country holding different expectations for their vehicles. Hence, it is important to consider those cultural differences that may influence the selection of an ethical setting. [27] explains that people's behaviour may not always match their moral judgements; it is possible that one person judges some action to be morally correct but decides to take a different course of action. In their experiment, they compare responses from British and Chinese nationals about the classic trolley problem. They found that Chinese participants were less willing to take action than the British, giving reasons like fate and respect for human life.

### 2.1.5 Proposed Implementations

Although there has been a lot of research in accident management, few of these studies address the moral questions of decision-making, aiming mainly at collision mitigation or avoidance solutions. However, as already mentioned, it is important to know the reasons behind these decisions so that they are defensible in a legal and ethical framework.

An ethical framework alternative to the utilitarianism, based on Rawls' moral theory is presented on [28]. The idea behind it is to achieve a Pareto-optimal solution; this means a solution where no participant situation can be improved without making another participant situation worse. The algorithm is based on the idea that the participants do not know what role they are playing (passenger, pedestrian, etc.) hence it is in their best interest to maximise the utility of the worst outcome. To make a decision, the algorithm takes into account the probability of survival of each participant on each possible manoeuvre. Within this set, the algorithm selects the actions that have the lowest payoff and a new set is created. From this new set it is selected the manoeuvres that have the highest payoff. This process is repeated until just one manoeuvre remains, this would be the decision output. If two or more actions tie with the same pay off, then the decision is randomised.

A three levels system to be implemented as the technology becomes available, is presented on [2]. In the first phase, a rule-based system would be encoded, these rules should be agreed on by ethicist, automakers and lawyers. The behaviours selected would be those that minimise damage. Although a problem with this implementation is that, in case of a scenario not covered by the rules, the car would just brake and evade, which in a collision situation, does not provide a real solution. Nevertheless, in phase two, machine learning techniques, like neural networks, would be implemented to increase the vehicle understanding on ethical decision. A shortcoming with this approach is that, if the neural network is not trained with a diverse and appropriated set of scenarios, the morals learnt by the algorithm can show extremist behaviours or discriminatory biases. Phase three of Goodall's approach consists of receiving feedback from the automated systems to understand the logic used to make a decision. However the implementation of this step may be slow as more research is needed on how to extract such information from neural networks.

These previously presented implementations are to be pre-programmed in the AV. Alternatively, Contissa et al [29] presents a customisable knob with three broad settings: (1) an altruistic mode which gives preference to other people lives; (2) an impartial mode, where passengers and other persons have equal importance; and (3) an egoistic mode, where the passengers lives have preference. The author also argues for a continuous mode, in which the knob would allow the user to select the weight of the passenger life relative to other persons. The system would take into account the probability for the passenger and third parties to suffer harm resulting from the AVs decision, to select the option with the smallest disutility. A problem of this implementation is that selecting too egoistical value would cause the vehicle to protect the passenger at all costs even if the probability of harm is very low, unnecessarily exposing the life of the pedestrians. The authors argue that the limits over the value that the passenger gives to their life should be regulated by law to avoid this situation.

## 2.2 Legal Considerations

One of the most attractive benefits of the introduction of AVs to our roads is that they may help to reduce the number of traffic accidents. Although thousands of lives may be saved, it is important to recognise that fatalities cannot be completely avoided. The intervention of AVs would create a new set of victims that in other circumstances may have not been damaged [30]. This situation creates a trade-off between the lives that can be saved and the lives that would be affected by AVs. Even if it can be said that the greater good should prevail, this is a trade-off that may not be sustainable from a legal perspective. However some researchers argue that this not may be the correct way to interpret the situation, as the introduction of AVs would reduce the chance of an accident and that is to the benefit of everyone, regardless of whether an accident occurs or not [31].

With autonomous cars already being tested in different cities, the first accidents have already happened [32, 33]. However, there is still no legislation in place to determine who is responsible for the damages caused by AVs. It is important that these regulations are in place before these vehicles are available to the public [34]. There are three main legal branches that cover AVs [35]:

- *Administrative law* oversees issues such as the traffic rules and technical norms for the operation of AVs on public roads;
- *Civil law* covers civil liability for injury or damage and product liability related to damages caused by a defective product;
- *Criminal law* manages responsibility when a crime has been committed.

One of the questions that researchers have been exploring relates to who would be responsible for the actions of an autonomous vehicle: owners, manufactures, programmers or the machine itself. Some authors suggest that the owner of the vehicle should be considered morally responsible in case of an accident [31]. This perspective could be considered as a shared responsibility within all the owners of AVs, since they are introducing a risk by using this type of vehicles. This can be applied as some sort of special insurance, tax or collective fund to compensate those affected [36], [37].

There are also proposals to protect the automakers from too harsh legislation, that could steer them away from the development of AVs. To limit the criminal responsibility of the manufacturer, a margin of tolerance for errors that may occur from the programming of the vehicles could be set [36]. Moreover, sacrificing the passenger to protect somebody else can be acceptable assuming that the passenger has signed a contract stating in clear terms that she or he understands and accepts this decision [38].

Under current legislation, some of the principles that can be applied to the programming of AVs are as follows. The doctrine of necessity [38] is a legal approach in criminal law to regulate those cases where damage has been intentionally inflicted to a third party when avoiding all evils is impossible. Within this doctrine, the theory of justification is found. It dictates that under extraordinary circumstances an otherwise prohibited action can be accepted. Programming a vehicle to harm under certain circumstances (to target someone that anyway was going to be harmed, in order to protect more people) can be acceptable as part of the theory of justification. However, to intentionally harm a third party unaware of the situation (swerve and harm a passer-by in the sidewalk) would not be acceptable. Nevertheless human lives should always be protected over property [36]. Some authors hold that ethical egoism should not be programmed in the AVs as it cannot be viewed as legally correct [39]. However utilitarianism also poses a challenge to the legal system since, in a democratic system, everyone has rights that cannot be ignored to justify social benefit.

Nevertheless, even if some guidelines can be drawn from current legislation, scholars suggest that governing agencies should review their legislation incorporating feedback from the public, the industry and relevant government agencies [40, 41, 35, 42]. Including the views of all the affected parties would help to establish clear rules to help boost the development of intelligent vehicles.

## 2.3 Advanced Driver-assistance Systems (ADAS)

Different approaches have been proposed for the development of automated systems. The earliest methods to support lane change assistance are in the form of warnings delivered to the driver through visual channels to aid them in the decision-making. Further developments perform the lane changes in an automated way on constrained scenarios and more autonomous systems can navigate in a highly automated fashion.

In general, to achieve automated driving certain tasks need to be completed. A mapping of the environment along with a priori knowledge such as the road type or car status are the prime source of information to define a suitable path to follow. Algorithms for path planning and collision avoidance are then fed with this information to derive secure manoeuvres. Once the system has defined possible actions to perform, a decision-making module is needed to select the best option. Finally, a control module is used to execute the preferred manoeuvre through the actuators in the vehicle. Even though all these aspects are important for the development of autonomous vehicles, this literature review is focused in the decision-making stage, as the scope of this is to understand how this system decides what action to perform. In the next sections, a review of some of the systems found in the literature is provided.

### 2.3.1 Warning systems

Different methods have been used to aid the decision-making problem, the simplest of them just sends a warning to the drivers, usually information about lane change, to aid them to reach a decision. Bayesian networks are used in [43] and [44] to derive a preferred manoeuvre. These are formed by chance nodes, decision nodes and utility nodes. The chance nodes represent the different situation variables that are determined from the perception results. The estimation is in terms of if it is possible, impossible or safe to perform a change lane. This probability is used along with a utility table to evaluate the most appropriated manoeuvre. The utility is assessed in terms of safety and it is obtained from a conditional utility table. This table is constructed considering all the combinations of the situation (lane change left, lane change right, ego lane) and the decision alternatives (keep lane, change left, change right). The resulting utility depends on the current situation and the decision. The table is constructed on the basis that a human driver is behind the wheel and considering German traffic laws. The decision made by the system is presented to the driver in the form of a visual warning. The final decision is based on the expected utility, the higher the utility the more preferred.

Other warning systems like the one in [45], performs a situation assessment and derives a warning using ensemble learning methods. The authors state that, a higher classification accuracy in the situation assessment and decision-making can be achieved using these methods. The methods used are random forest and AdaBoost. The output of these methods is a binary classification, “lane keep” or “lane change”. The model was built with observations of traffic data from segments of two USA highways, speed and position were obtained through observation of 1-s intervals and from there two driver behaviours were identified, either lane changing or lane keeping. This data was used to compare the output of the model. Factors such as vehicles speed, speed difference and distance between vehicles, were identified as decisive to change or keep lane and were used as inputs of the model.

### 2.3.2 Automated lane change

More advanced lane assistant systems than warning systems, evaluate the situation and perform the appropriate manoeuvre in an autonomous way. A state space model is used on [46] to assess the situation and derive a behaviour decision. With information from the sensors, the road is divided into 8 segments, three front areas, two side and three rear areas. The state variables provided by the object tracking system like position, velocities, dimensions of the tracked vehicles and the time to collision, are stored into a state matrix that is used to obtain the traffic situation. The process that leads to a manoeuvre is structured in a hierarchical fashion. At the top is the situation evaluation, next is the driving request (keep desired velocity, change lane, keep lane, etc.) and below are the feasibility assessment and the execution. If a lane change is requested, a Lane Change Gap Approach algorithm is used. This assesses the feasibility to perform a lane change, taking into account the gaps on the target lane and the trade-off between time and safety. The implementation of the decision-making system is through an event-triggered deterministic finite state machine that includes the longitudinal and lateral states, modelled using hierarchic abstraction levels.

A system to assess if a lane change manoeuvre is feasible and desired, as well as when is the best moment to perform it is presented in [47]. The decision-making process is done through a utility function representing the utility of each available lane. To perform a safe change of lane, the vehicle should maintain a safe distance to all surrounding vehicles while positioning itself in a gap between vehicles in the target lane. The trajectory planning is computed as longitudinal and lateral Model Predictive Control (MPC) problems. To perform a change lane manoeuvre on top of being feasible it also needs to be desirable. Uncertainties from the sensor readings are handled increasing the safety distance that the vehicle must maintain to others. For a lane change, the expected utility of the desired lane should be higher than the current lane. Once that a lane change manoeuvre is assessed desirable, the algorithm determines an appropriate traffic gap in the target lane in terms of the maximum and minimal position that the vehicle can reach at a certain time instance, the minimum safety margin and the position of other vehicles. To plan a safe and comfortable lateral and longitudinal trajectory, constraints are applied to position, velocity and acceleration.

### 2.3.3 Vehicle navigation

Other approaches study more specific urban situations like what to do when the vehicle encounters an uncontrolled intersection or a narrow road, [48] is a decision-making system presented to assist in this cases. Knowledge about the car and the environment is represented by ontologies. A conceptual description is used for modelling data, it describes objects trough statement in the form of subject-property-object. For example, the road would be the subject, and it has properties like intersections, lanes, road segments, max speed, etc. The object is used to assert the relations between lanes as well as to identify the driving direction. Two ontologies are used; the first is the map ontology that contains information about driving environment information like, different types of roads, intersections, lanes, and the relations among them. Second ontology is the control ontology, used to represent paths and driving directions of the vehicles. Using properties of control, the path segment, that can be any part of the road like intersections or lanes, is indexed and is defined by start and end nodes, also it can be linked to the next connected path segment. Another property present is the collision warning defined to indicate upcoming collisions. When a collision warning is detected, the rules reasoner is executed according to the situation to give as result an action to be performed.

Fuzzy theory is also applied for decision-making, [49] presents the behaviour decision module using a vision system called Expectation-based Multifocal Saccadic vision (EMS-vision) system. Decisions are processed in three units; the central decision unit has the highest authority for decision-making and oversees all the resources available to fulfil the goal. Also solves conflicts between the other two modules. Behaviour decision for locomotion in charge of the locomotion capabilities available from the locomotion expert vehicle control. Behaviour Decision for Gaze and Attention controls the vision strategies to satisfy the needs of the perception expert. The experts refer to the data processing and knowledge needed to control submodules able to perform certain capabilities like follow a lane or perceive obstacles in the road. These capabilities are represented in a capability net which is used to communicate the behavioural decision unit with the experts through commands to control the capabilities such as initialize, parametrise or stop. The situation analysis module, part of the central decision unit, performs the calculation and storage of linguistic variables that represent the situation aspects. These variables are evaluated by fuzzy IF-THEN rules and suitable capabilities and their parameters are defined. In the capability control module, the rules that are mostly fulfilled trigger their respective capabilities, if a rule is no longer fulfilled then the capabilities associated with it are stopped or deleted.

In [50] a navigation algorithm based on fuzzy rules is presented. The algorithm first analyses the environment to detect any dynamic obstacles and then estimates the state vectors, then by means of fuzzy rules the velocity for the robot is defined. The optimal velocity would allow the robot to decelerate to wait for the object to pass or to accelerate in order to avoid the object. The decision-making process is started if, after estimating the path of the obstacle, this interferes with the robot path. The objectives are to safely reach the target point in the smallest time possible. Two fuzzy goals are implemented, which are defined in terms of time that the robot and the object takes to arrive to certain point, and a control parameter defined by the designer. An exponential sigmoid function is used to design the membership functions of the goals; these can be expressed with a min-max operation where the alternative with the highest value is selected as the velocity to implement.

### 2.3.4 Naturalistic driving

Other researchers have focused their work on analysing how human drivers make decisions, especially in urban scenarios. In [51] the authors constructed a virtual traffic environment to collect data from non-professional drivers. The collected information was divided into two categories, the first involving the vehicle movement parameters such as the vehicle velocity, acceleration and steering wheel angle. The second was basic information about the driver like age and driving experience. The experiment consisted on a transition from car following to lane changing. After analysing the obtained data, it was stablished that in car-following conditions the vehicle maintains a lateral acceleration in the range of $-2\ to\ 2\ m/s^2$ hence a lane changing manoeuvre starts when the acceleration exceeds this threshold. To extract driving decision rules, the vehicle data for lane-change is discretised through a break point on each attribute (distance, velocity and acceleration) to construct a decision table, taking the lateral acceleration as the decision attribute. The decision rules generated are extracted from the decision table and are based on the lateral acceleration and the relative distance between vehicles. An analysis of the driver's behaviour showed that, the relative speed and distance between vehicles are the factors that have more influence on the decision-making once they exceed or fall below a certain threshold. With this information, a car-following model based on driver's cognitions was derived. This was used in the tactical planning module, which chooses the driving strategy of the vehicle.

### 2.3.5 Fully autonomous vehicles

Petri nets and multiple criteria decision-making are used on [52] and [53] to allow fully autonomous driving. The decision-making module is divided in two stages. In the first, all the available manoeuvres are evaluated and only the safe ones are selected. In the second stage, the selected safe manoeuvres are evaluated using the additive weighting method to choose the most appropriated one and the attributes to execute it. The decision-making is processed through a hierarchical structure, where different levels of objectives are specified. These objectives can be measured by their attributes. The attributes serve the purpose of rank the objectives, this is done assigning weights to the attributes. Each possible manoeuvre can be executed with different parameters. When discrete values are assigned to these parameters, they represent the decision variables in multiple-criteria decision-making. An objective function is also implemented, this represents the level of achievement of an objective by an alternative. The efficiency of an alternative to fulfil the goal is measured through utility functions evaluated based on heuristics that represent the preference of the driver on a scale from zero to one. The weight assigned to each attribute varies depending on the type of road, as certain situations require different actions.

There are other approaches that do not have a decision-making module per se but derive decisions from other methods, [54] presents an optimization-based manoeuvres system that, takes into account the passenger comfort, safety distances and the dynamics of the vehicle. Data driven vehicle dynamics along with optimization-based manoeuvre planning are used to compute a collision free trajectory. The dynamics of the vehicle are modelled in a way such that all the states that can lead to skidding or loss of control are excluded. A state that includes constrains and a set of weights for the manoeuvre optimization function is used to bias the generation of a guiding path. A Boolean function is generated to determine if the set of controls is feasible or not. To select the best path, a cost function is minimised. This integrates a path cost that measures the vehicle success in following a route and the comfort cost that penalizes any control that may be uncomfortable for the passengers. Lane changes are often dangerous or in some cases prohibited by law, hence a manoeuvre cost is added to discourage this action. A proximity cost is also added to avoid passing too close to neighbouring objects. As a result of this stage, the most appropriate set of controls is chosen and passed to a PID controller that executes them.

### 2.3.6 Collision avoidance

A proposal to design decision systems for collision avoidance, considering a human driver, is presented in [55]. Every time that the system intervenes, whether with a warning or an action, it generates a cost for the driver. This generated cost is important because systems with high cost are detrimental not only for the comfort of the driver but also for safety, as they can divert the driver attention. The proposed process aims to increase safety and decrease cost. These two objectives are labelled as accuracy and liability respectively. In the first instance, the driving state is determined through the parameter Time to Lane Crossing, which is the parameter that drivers use to estimate when a steering action is needed. With an estimate of this time, the unsafe driving situations are determined. Next, a set of possible actions to be carried out by the system is created. The decision problem is then to define which of these actions to select. The accuracy (utility) and liability (cost) attributes are the utilities used, the utility of an action is determined by the probability that the action causes consequences and the accumulated safety payoff caused by the action. The cost of an action depends on the probability that the action interferes with driver autonomy and the accumulated cost incurred by these actions. To select an action, two principles are implemented, the burden of proof and the domination principle. In the first, the safety enhancement and the incurred cost are evaluated, the satisfying set is created by those actions whose accuracy is high enough to justify their cost. Then a dominance analysis is performed over this set yielding the strongly satisfying set that consists of the non-dominated actions. A driver parameter is also implemented, this will further aid in selecting the best action, this parameter can be adapted to different drivers, as some may be comfortable with continuous warnings to enhance safety and others may prefer a less invasive system.

A Collision Mitigation system that uses different levels to allow a flexible parametrization, is presented in [56]. This work is explained in more detail in the following sections, as it has been used as a base for the future work of the present research.

### 2.3.7 Existing Proposals

Dennis et al [57] propose a framework where the ethical decisions are drawn from the ethical codes and regulations governing the profession related to the function of the machine. They assume that each professional domain has already developed ethical principles and substantive rules to evaluate how ethical an action is when there is no ethical option available: the system should then take the least unethical decision. The authors provide an example of an autonomous aircraft, considering that the machine should act as a pilot would do. In their proposal they establish a set of ethical concerns and define which of them are more ethical to violate. The system presents two operational modes. The first is controlled by the pre-programmed plans where the programmer assumes responsibility and the second mode where the ethical reasoning is needed to operate when no plan is available or all plans have already been implemented but failed. A rational agent determines which of the new plans, supplied by an external planning mechanism, are the most ethical to follow.

A benefit of this system is that it can evaluate how ethical a decision is but is not limited to only perform ethical actions; an unethical action can be performed provided that there are no more ethical actions available. Additionally, the framework is designed to be verifiable, meaning that if a decision made is considered unethical it can be proven that the system believed that it was the minimally unethical action from those available. A limitation in a system like this is that driving a regular vehicle is different to operating an aircraft. While there are traffic codes that must be followed, these vary from country to country and do not always resolve ethical issues.

Dennis et al's concept - defining a set of ethical concerns and arranging them in order of importance according to an ethical policy, used in [57] serves as the basis for our proposal presented in Chapter 3.

More recent works like DeMoura et al [58] propose the implementation of three policies based on the ethical principles of contrarianism, utilitarianism and egalitarianism. The decision-making algorithm that they propose is based on a Markov Decision Process (MDP) which controls the vehicle in normal conditions. This is done with the use of a reward function based on the vehicle performance, measured as the distance to the endpoint of the path being followed. The action consequences are measured by the adherence to the traffic code and proximity to other road users. The traffic rules implemented in this model are a velocity limit and the presence of other vehicles on the opposite lane and the pavement. The reward function determines the best trajectory to follow without involving ethical considerations in this process. Ethical considerations are introduced through the estimation of the expected harm for each road user, this value is applied to the ethical policies selected. For the contrarian approach, their objective is to minimise the larger expected harm. For the utilitarian part, the objective is to minimise the total amount of expected harm, to find the least wrongful action. In the egalitarian approach, they distribute the expected harm among the road users to avoid large discrepancies. In the egalitarian decision making, the number of passengers or pedestrians is not considered.

In Evans et al [59] the authors propose a strategy called Ethical Valence Theory (EVT), in which they indicate that each road user holds a different moral claim on the vehicle behaviour and has certain expectation of how they should be treated by the AV. This philosophical approach is known as a form of moral claim mitigation where the objective is to find an optimal response to the claims made by the road users. It is assumed that the strength of this claims can vary according to the type of road user; a pedestrian claim to safety is stronger than that of a passenger as it could be more seriously injured. The vehicle control implemented in this approach is the MDP control presented in [58] who estimate the expected harm in the same way. For the ethical approach they define two sets of rules, the first one covers the interaction between road users and the second set reflects the traffic code. In the first instance, each action available is assessed based on the two sets of rules defined. If an acceptable action is not possible, the EVT chooses which action to execute. The valence concept is introduced here, where the valence is the degree of social acceptability that is attached to the claim of the road users. The strength of the valence depends on physical characteristics that are ranked according to their importance and vulnerability. For example they use the age and type of road user, so the valence of young pedestrians is more important. The valence value and the expected harm are used in the ethical deliberation to select an action that maximises the welfare of all road users.

### 2.3.8 Conclusion

After an initial review of the presented literature, certain gaps were identified; the most critical is the lack of analysis of the system's behaviour in case of an unavoidable collision. All the reviewed systems are tested under very particular and staged conditions, however real-world driving situations, especially in urban environments, are vastly more complex.

For example, in [53], if there is no safe manoeuvre the vehicle can perform, it stops and waits until one becomes safe. However in a real-world scenario, this kind of actions may not always be safe.

Additionally, since the presented systems do not consider the possibility of a collision, an ethical discussion about the decision-making process is not present in the literature. Ethical considerations, according to German and US government guidelines, is an essential part in the development and future roll out of autonomous technology [60], [61]. However little research in the matter has been done leaving a gap that needs to be fulfilled.

The novelty of our work is that, rather than assuming a fixed ethical policy, we cater for a multiplicity of theories (five in the present work) and allow a theory to be selected by the user. This approach results in a different manoeuvre depending on the theory that is selected by the user, accommodating for the ethical views and preferences of each individual, while at the same time upholding pre-established rules and laws as implied by [29]. In the proposal by Dennis et al the ethical concerns are defined by a specific set of rules to follow and a decision is made based on how unethical the rule is to break and does not differentiate the gravity with which each principle is violated, simply establishes that violating two separate principles is worse than violating just one. The system proposed in the present work does not contemplate rules. The ethical concerns are defined by considering all the possible participants in an accident scenario. These concerns are then prioritised according to what is expected in each ethical approach.

There has been a recent rise in research about ethical decision-making for autonomous vehicles. The work presented by DeMoura et al [58] presents similarities to our research. They present a system capable of deriving an action to be implemented based on the evaluation of different ethical theories. However our system is different because we have defined eight ethical concerns and all of them relate to a different type of road user (inside and outside of the vehicle) with varying degrees of importance, while they do not differentiate between the different types of road users when applying their ethical settings.

Comparing our work to the one presented in Evans et al [59] we can note two differences. First, they are using physical characteristics like age to determine how acceptable is to harm a road user, which could lead to potential discrimination. In our system we do not include any differentiation between people. Another point to note is that they assume that the system has a lot of knowledge about the road users, and while the type of user (pedestrian, cyclist, etc.) and possibly their age can be identified by vision systems and object classification algorithms, information about the passengers of other vehicles might require communication capabilities, from which privacy concerns could arise.

Another difference is that our system aims at allowing the user to select the ethical theory that they prefer, instead of using a predetermined one.

# Chapter 3

# Ethical Perspectives

In this chapter we propose a novel approach for a decision-making system for autonomous vehicles, followed by the identified ethical concerns, their definition and analysis of these according to each ethical theory. Finally a discussion about some of the issues that could arise with this implementation are touched upon.

## 3.1 Ethical decision-making System

We propose a system for fully autonomous vehicles, capable of taking decisions based on the user's preferences. The user will be able to select an ethical view that aligns with their own through a user interface in the vehicle's screen. This setting will be stored in the vehicle's memory and will be retrieved if needed. Through the normal operation of the vehicle, if a dangerous situation is encountered, the vehicle will calculate the possible manoeuvres that it can perform. If a safe manoeuvre is available this will be implemented. In case that there is no possibility to avoid a collision then the vehicle will evaluate the possible manoeuvres against the user preferences and the legal requirements previously stored in the vehicle's memory and will perform the one that complies with the greatest number of the constraints. To the best of our knowledge, a system like the one presented here has not been studied before.

The system's proposed aim is to give the user the opportunity to have an input on the behaviour of the autonomous vehicle. While the vehicle computer will be the one making the final decisions, it will consider the wishes of the user. With this, we want to give a different perspective over the issue of who should make the decisions in an autonomous vehicle. Our system would accommodate cultural and ethical preferences. And also address the discomfort some users experience when a machine takes the decision. In this chapter we present an analysis of the ethical theories that will be implemented in our system and the different ethical concerns that will form the base guidelines to establish the course of action.

## 3.2 Ethical Concerns

Similar to the work of [57], we have defined eight ethical concerns for the vehicle to follow. These concerns were defined considering the most likely participants in an automotive accident; people, vehicles, animals and objects. Current ethical discussions and governmental policies for the development of AVs [62, 60, 16] were also taken into account. A distinction between passengers and people outside the vehicle is made to reflect situations where it is not possible to protect everyone and an ethical decision is required. To be able to evaluate the same concerns in all ethical theories, these have been extended to fit in the context of AVs The concerns are as follows:

- Harm the least possible number of people (c1)
- Do not harm passengers (c2)
- Do not harm people outside the vehicle (c3)
- Inflict the least damage possible to people (c4)
- Do not harm vehicles with passengers (c5)
- Do not harm children or incapacitated people inside the vehicle (c6)
- Do not harm animals (c7)
- Do not collide with inanimate objects (c8)

As discussed in Section 2.1.1, we have selected five ethical perspectives to be applied to the system. These are utilitarianism, distributive justice, Kantian ethics, altruism and egoism.

The ethical concerns have been defined in terms of harm to self and others; depending on the selected ethical approach, these will have more or less importance to the system. As

Table 3.1: Ethical concerns ranked according to each ethical theory from highest priority at the top to lowest priority at the bottom

| Ethical concerns (for reference, unranked) | Utilitarianism | Distributive justice | Kantian | Altruism | Ethical Egoism |
|---|---|---|---|---|---|
| Harm the least possible number of people (c1) | c1 | c4 | c3 | c3 | c2 |
| Do not harm passengers (c2) | c4 | c1 | c1 | c2 | c3 |
| Do not harm people outside the vehicle (c3) | c2 | c2 | c2 | c1 | c1 |
| Inflict the least damage possible to people (c4) | c3 | c3 | c4 | c4 | c4 |
| Do not harm vehicles with passengers (c5) | c5 | c5 | c5 | c5 | c5 |
| Do not harm children or incapacitated people inside the vehicle (c6) | c6 | c6 | c6 | c6 | c6 |
| Do not harm animals (c7) | c7 | c7 | c7 | c7 | c7 |
| Do not collide with inanimate objects (c8) | c8 | c8 | c8 | c8 | c8 |

"passengers", we define any person that is travelling in the vehicle. The "people outside the vehicle" refers to either pedestrians or bystanders. A concern for "vehicles with passengers" has been added to make these vehicles distinct from our definition of "inanimate objects", where we refer to any element like empty cars, walls and poles. A concern related to children travelling inside the vehicle has been added. The distinction between harming the least possible number of people and inflicting the least damage possible to any individual is that in the first we are referring to overall numbers where fewer people damaged is better, and in the second it is permissible to harm more people if that means that the most damage to any individual would be less. For example, it would be better to harm two people but keep them alive, rather than sacrifice one to avoid harming the other. A distinction between animals and objects is included since animals are sentient beings that are important for humans, based on a recommendation from [16].

For the **utilitarian** approach, we are looking to maximise overall good, hence the most important thing is to try to harm the least possible number of people (c1) irrespective of whether they are inside or outside the vehicle. The next most important thing is to try to inflict the least possible damage to any individual (c4). Do not harm passengers (c2) and people outside the vehicle (c3) concerns are equally important as, in an utilitarian approach, their location is irrelevant.

When using the **distributive justice** principle, the most important thing is to inflict the least damage possible to people (c4) followed by to harm the least possible number of people (c1). In this approach it is also irrelevant if the persons involved are inside or outside the vehicle, hence concerns two (c2) and three (c3) are equally important.

In the **Kantian** view, the most important thing is to protect people outside the vehicle (c3), as we give them value as a person and avoid using them as means to our objectives. Next the system will aim to harm the least possible number of people (c1) followed by not harming the passengers (c2), and lastly to inflict the least damage possible to people (c4).

**Altruism** and **egoism** give a similar order to the concerns with the exception of the two first. For altruism, the most important thing is to protect the people outside the vehicle (c3), while for egoism, the most important thing is to protect the passengers (c2). One important thing to note here is that the egoism condition can only be applied when the number of people outside the vehicle is the same as the passengers, as it is not possible to value the life of one person more than the lives of two or more.

The order that we have given to these concerns, based on the different ethical approaches, is presented in Table 3.1. In first instance, concerns seven (c7) and eight (c8) always occupy the last two places, as animal and objects are always less important than people. Concern six, do not harm children or incapacitated people inside the vehicle (c6) is always above concern seven (c7). Concern five, do not harm vehicles with passengers (c5), is above number six.

### 3.2.1 Discussion

The system proposed in this work aims to allow the users to select an ethical behaviour that aligns with their preferences. A potential disadvantage of a system like this is that it could be misused. However, assuming certain conditions, this can be prevented. Examples of these conditions are: (1) the user will not have control over the driving at any time, meaning that any accident would be due to a vehicle malfunction or other external factor and not because the user acted in bad faith; (2) the system should be parametrised according to the law and this can vary for different countries. For example, trading off the lives of two or more people to save one is not permitted.

An argument against allowing the user to select the ethical behaviour of the system is that it could defeat the purpose of an autonomous system. However, while it is certain that the user acquires a duty and accepts the risk of using an autonomous vehicle, it is also true that external factors completely out of their hands can cause an accident and they still should have the right to decide how they want their vehicle to behave. This does not mean that they should not be held responsible for their decisions but rather to change the focus from a potential criminal liability to a civil one, as discussed in Section 2.2.

In legal terms, an issue that arises with the pre-programming of actions that may lead to people being harmed is if that could qualify as premeditation. Under current legal standards, planning an action that leads to someone being hurt is a crime. However, in the case of the proposed system, neither the programmer or the user knows the circumstances nor are expecting or planning for an accident to happen and since they do not drive the vehicle they can not purposefully cause an accident. Here, the user is simply setting behaviour parameters in case that something occurs. In the current legal framework of most countries, the use of autonomous vehicles has not been contemplated and it is hard to judge such vehicles under the same rules that apply to a human driver. New legal standards for these vehicles are under development. Hence, in the future, a legal framework that contemplates a system like the one proposed here could be possible.

Another question that arises is why not let a body of experts in ethics decide how these vehicles should be programmed. Although this approach could be a solution, the general public may not agree with the experts' conclusions. Examples of different ethical views can be found in many areas of contemporary life, such as abortion and immigration. Debate in these areas has existed for a long time, and even laws have been established and applied to some countries but not in others. These continuing debates suggest that reaching a consensus over these ethical questions is not straightforward and cannot always be universal.

## 3.3 Conclusions

In this chapter we have identified a set of ethical perspectives that can be applied to an autonomous vehicle. Ethical concerns for the usage of this vehicles have been identified and evaluated against the corresponding ethical theories and ordered according their importance in each particular theory.

A novel approach for decision-making systems has been presented. One of the advantages of this proposal is that it can be used to fit a wide range of user expectations about the behaviour of the AVs. A system with this flexibility can be attractive for both industry and users, allowing further development and acceptance of these vehicles.

# Chapter 4

# Decision-making system

The structure of this chapter is as follows. In section 4.1 we introduce our proposed system. In Section 4.2 we present the agent architecture. In Section 4.4 we give an introduction to the language used to develop the agent. In Section 4.5 we explain the logic for each ethical setting. Finally, in Section 4.6 we provide conclusions on the development of the Ethical decision-making agent.

## 4.1 Proposed System

In this chapter we derive a system that allows the driver to select the ethical settings defined in Chapter 3. Such a system will operate the behaviour of the vehicle in case of an unavoidable accident and will provide a new way of answering ethical questions about autonomous vehicles.

The part that differentiates this work from other avoidance collision systems is that we propose the inclusion of a moral component that reflects the user's ethical preferences. Unlike systems that develop moral standards through artificial intelligence, our system would give the user the power to decide which ethical vision suits their preferences best. This of course should be within a legal framework. At the time of writing this thesis, legislation for the usage of autonomous vehicles is not yet clear. However, different governmental authorities have published guidelines for the future operation of this technology [16, 61]. In this research, we will use this as a guide to tailor our system. To the best of the author's knowledge, this is the first system that aims to have input from the user on the ethical behaviour of the vehicle, rather than a unique behaviour setting decided by the programmers.

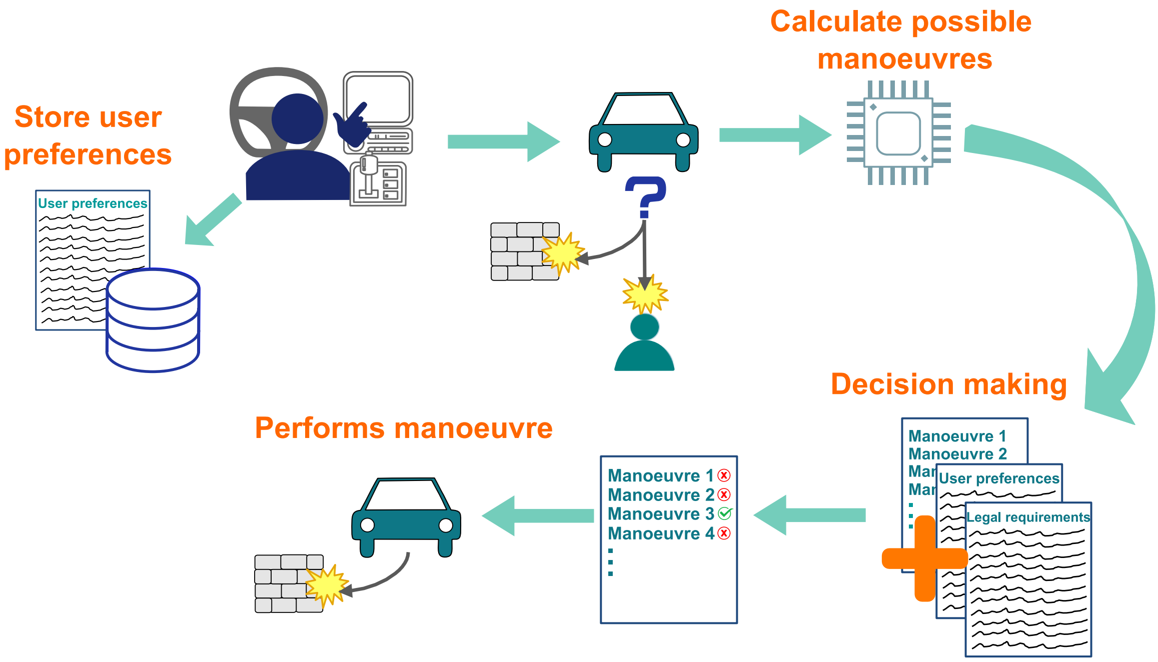

Figure 4.1: Proposed system.

The proposed system is presented in Figure 4.1. As a first step, the user will be able to select the desired ethical setting through a user interface accessible from the screen in the vehicle's dashboard. This information will be stored in the vehicle memory and retrieved if necessary i.e. during an unavoidable collision. When the vehicle encounters an emergency, it evaluates the situation and calculates possible manoeuvres. If the collision cannot be avoided it starts the decision-making phase. Here the user preferences are retrieved and the existing legal requirements are taken into account. Finally, a manoeuvre that complies with both conditions is selected and performed.

Different elements that influence the outcome of a situation can be considered to calculate the different scenarios, like the number and location of the passengers, the safety systems integrated into the vehicle, the objects in the vicinity and the type of collision; the system will process this information as illustrated in Figure 4.2. While the vehicle is moving, it will check for possible collisions. If a possible collision is detected, the system calculates the possible manoeuvres to avoid the collision. If the collision can be avoided, it performs the manoeuvre. Otherwise, it re-calculates the possible movements and evaluates them against the ethical settings selected by the user and any legal requirements of the vehicle. Once a suitable action is found, it proceeds to perform it.

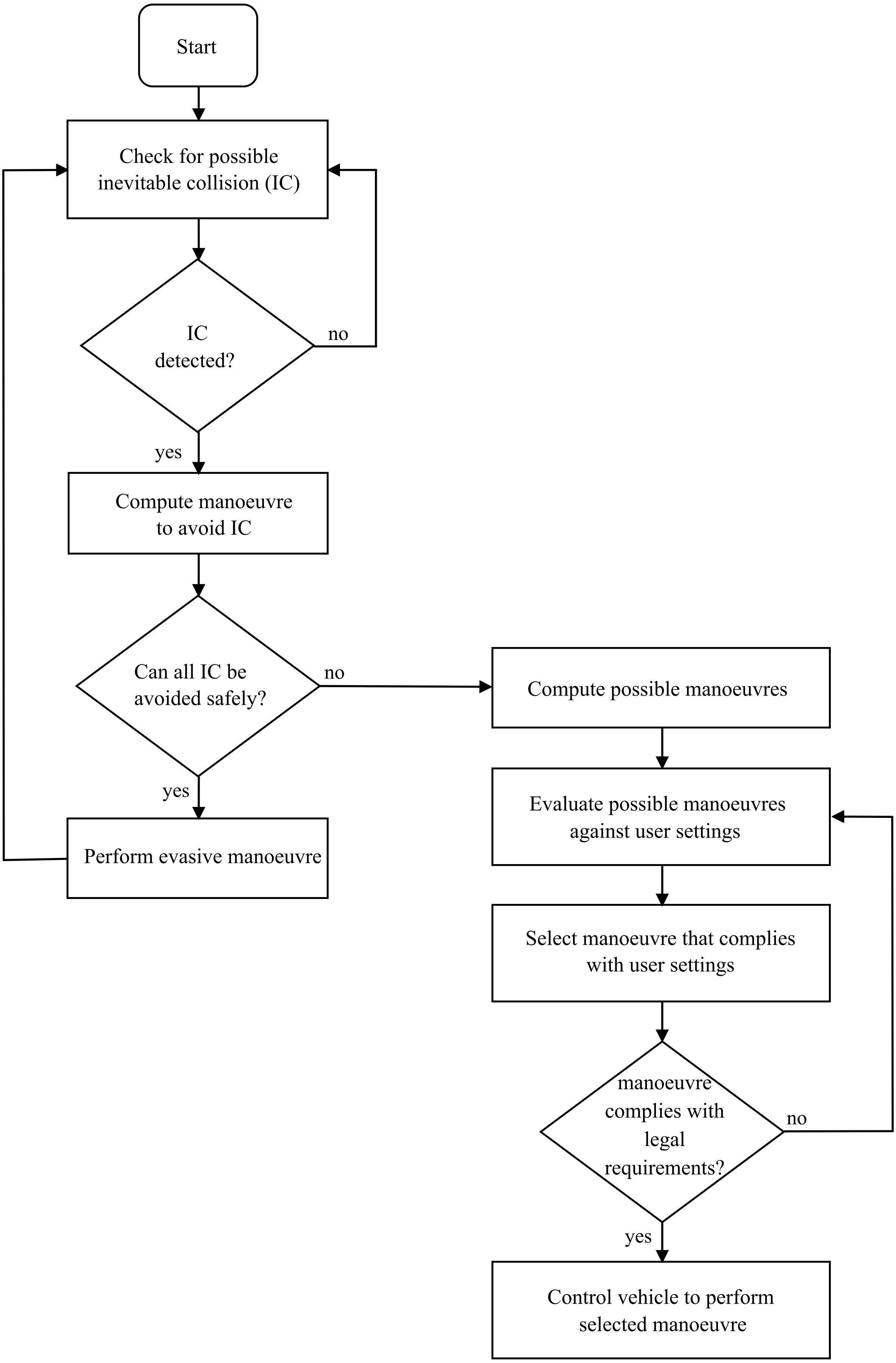
Start
Check for possible inevitable collision (IC)
IC detected?
no
yes
Compute manoeuvre to avoid IC
Can all IC be avoided safely?
no
yes
Perform evasive manoeuvre
Compute possible manoeuvres
Evaluate possible manoeuvres against user settings
Select manoeuvre that complies with user settings
manoeuvre complies with legal requirements?
no
yes
Control vehicle to perform selected manoeuvre

Figure 4.2: Flow diagram of the proposed system.

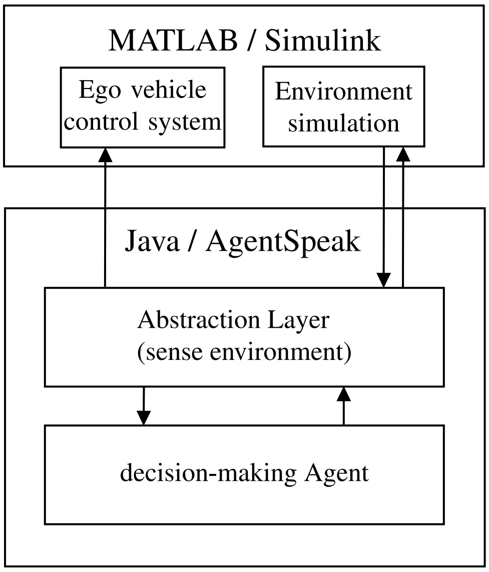

Figure 4.3: Architecture of the ethical agent and simulation

## 4.2 Agent Architecture

For the implementation of the system we selected an hybrid architecture similar to the one presented in [63]. Hybrid systems are defined as those that combine an agent-based decision maker with continuous control systems [64]. In our case, the ethical decision-making agent is implemented in Jason; the vehicle control system and the environment are simulated in Matlab, as shown in Figure 4.3. Jason is a Java interpreter to develop BDI agents. This agent will be in charge of perform the ethical the decision-making process. The vehicle control and dynamics are implemented in MATLAB along with the sensing and the physical environment. Between the agent and MATLAB, there is a layer in Java that processes the information sent between them. This abstraction layer translates the sensors readings to perceptions that are used by the agent as its source of knowledge about the word. This layer also translates the instructions of the agent to be implemented in the vehicle control.

## 4.3 BDI agents

The *belief-desire-intention* (BDI) model is based on the theory of *practical reasoning* presented by Michael Bratman in [65]. This theory involves deciding the *goals* the agent wants to achieve (-*deliberation*) and *how* it is going to achieve them (-*means-ends reasoning*) [66]. This process starts by understanding the options available, these options generate a set of alternatives. From these alternatives, one must be chosen and commit to achieving it. The chosen options become *intentions*.

The three principal elements of the BDI model are:

- Beliefs: the information the agent has about the world.
- Desires: the possible affairs that the agent might want to accomplish.
- Intentions: the states of affairs that the agent has decided to work towards.

For the *deliberation* stage of the *practical reasoning*, intentions are central in the BDI-model. Intentions have some key properties; the first is that they lead to action, i.e. if the agent commits to an intention, is reasonable to expect that the agent will act to achieve such intention. Secondly, intentions persist, i.e. the agent needs to remain committed to its intention to achieve it. However, if it becomes clear that it is not possible to achieve what the agent is expecting or if the reason for having that intention disappears, the intention should be dropped. And lastly, intentions are also related to the agent's beliefs about the future. The agent must believe that it can achieve its intention, otherwise, it would be irrational to act to achieve something that it does not believe can be achieved.

An issue with the BDI model is that the agent needs to achieve a balance between the properties of the different intentions. In consequence, on occasions the agent needs to reconsider its intentions to check if the intention needs to be dropped either because it is already achieved or it believes that it will never be achieved. From [67], we get that for static environments a pro-active goal-directed behaviour is enough. However, for dynamic environments, a reactive agent is more necessary.

In the *means-ends reasoning* part, also called planning, the system takes a goal or intention, the current beliefs of the agent about the state of the environment and the actions available to the agent. With this data it generates a *plan*. This plan then is implemented and once it is completely executed the goal used as input will be achieved.

Additionally to the set of beliefs, desires and intentions, a BDI agent has other functions to aid in the process of *practical reasoning*. A *belief revision function* used to update the agent beliefs, and an *option generation function* that determines the options available based on its current beliefs and intentions. From this function we get a set of *current options*. There is also a *filter* function which represents the deliberation process and determines the agent's intentions based on its current beliefs and desires, a set of current *intentions* which are those affairs that it has committed to trying to achieve and an *action selection function* which determines an action to perform.

## 4.4 AgentSpeak and Jason

AgentSpeak is a programming language for rational agents based on logic programming for the development of BDI agents [68]. The main elements of the language are a set of beliefs and a set of plans. The beliefs of an agent represent the things that it knows about the world, usually through perception. Plans are actions that the agent can perform to achieve a desired state, they have the following structure:

```
triggering_event :  context <- body
```

- A `triggering_event` can be a change in the environment or the acquisition of a new goal, this event may commence the implementation of a plan. There are four types of triggering events: the addition of beliefs or goals, and the deletion of beliefs or goals.
- The `context` is a series of literals that must be a logical consequence of the agent's current beliefs for the plan to be applicable.
- The `body` of the plan is a series of actions to follow in order to handle the event that triggered the plan. These actions will not necessarily have an impact on the environment. They can be sub-goals or internal actions.

Jason is a Java-based interpreter for an extender version of AgentSpeak. This interpreter allows us to combine the structure of an agent written in AgentSpeak with calls to Java code. In Jason, the agents operate following a reasoning cycle illustrated in Figure 4.4 (reproduced from [1]).

The cycle the agents follow consists of ten main steps that are labelled with numbers. In rectangles, we find the main components that determine the agent state: the *belief base, plan library, set of events* and *the set of intentions*. The *check context, unify event* and *execute intention* functions, represented in circles, are essential parts of the interpreter. Diamonds are *selection functions*; these mean that they select an item from a list i.e. an event, an

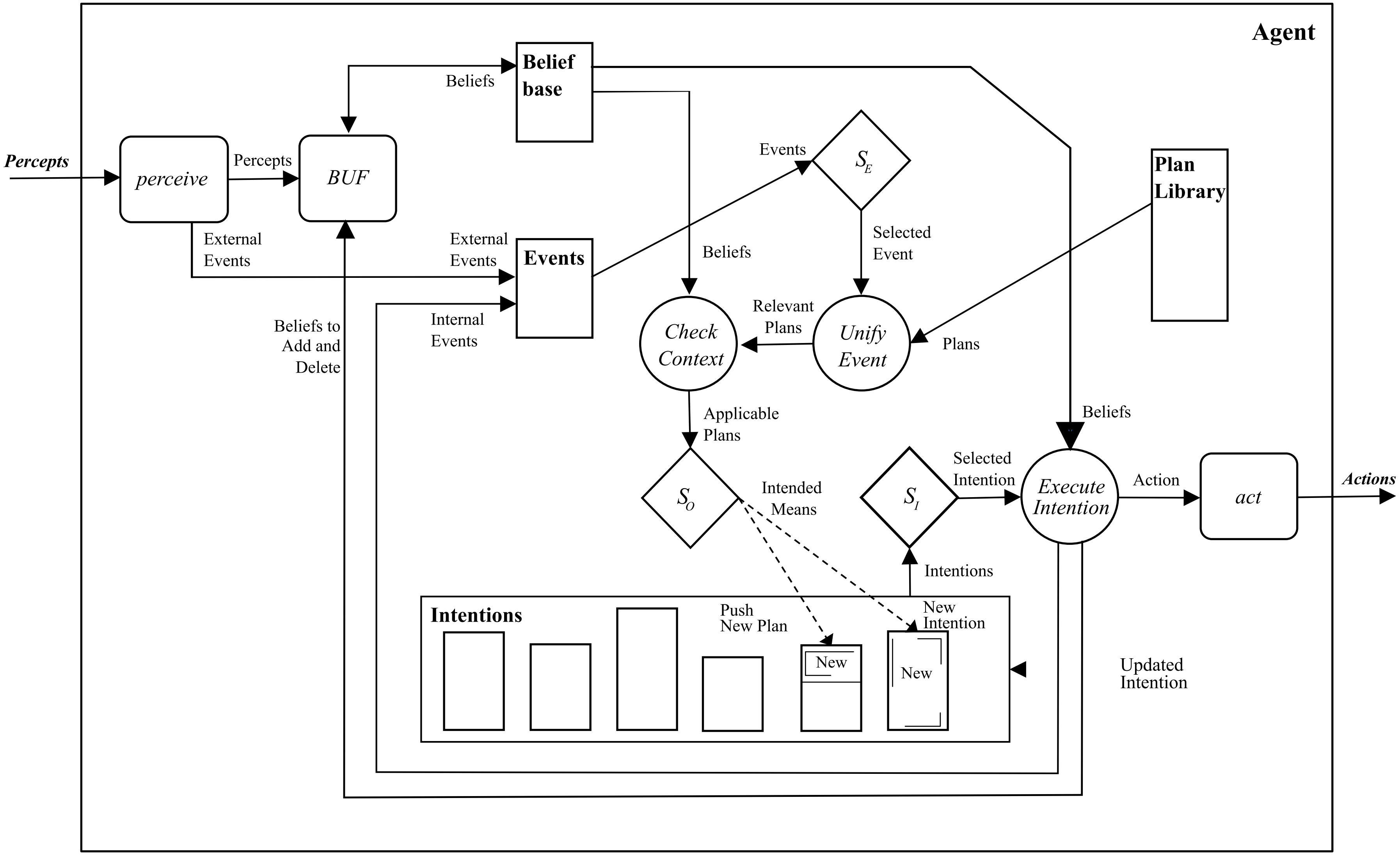

Figure 4.4: Jason Reasoning Cycle [1]

intention, or a plan. *Selection functions* can be customised if the agent needs to prioritise an specific event, intention or plan.

At the start of the program, the initial beliefs initialise the *belief base*. This process causes belief addition events that are incorporated to the initial *set of events* along with the initial goals. At this stage, the initial plans are added to the *plan library*. Every cycle starts with the perception of the environment, a list of percepts is compiled and send to the *belief update function (BUF)*. In this function, the list of perceptions is compared to the *belief base*. If a perception is not on the *belief base*, then it is included, and if a belief on the *belief base* is not present in the list of perceptions, it means that it is no longer perceptible in the environment and is deleted from the *belief base*.

The changes in the *belief base* generate new events. On each cycle, a single event is selected by the *event selection function* $S_E$. However, there might be more than one event that needs to be handled, so this function can be customised according to the priorities of the agent.

Once an event is selected, it is unified with the plans in the *plan library*. Then, the event is compared to the *triggering event* on each plan and a list of *relevant plans* is created. From this list, the context of each plan is checked and a list of *applicable plans* is obtained. From here, the *option selection function* $S_O$ selects a plan. Plans are ordered in the order in which they were written in the source code.

Next step in the cycle is to select an intention. This is done through the *select intention function* $S_I$. If the intention is a plan, the body of it is executed. If the intention is a *test goal*, the intention is updated, if is an *achievement goal*, the action is executed and this ends the cycle.

## 4.5 Ethical Decision-Making Agent

An overview of the decision-making agent system is presented in Figure 4.5 using the Prometheus notation [69]. The agent is in the centre and the stars represent the perceptions it receives. The possible perceptions that the agent can have are:

- `collision_Avoidance(action)`: When a collision is perceived, the agent is informed of the two possible actions that can be performed to avoid it: brake or steer.

- `PAX(No.)`: The number of passengers inside the ego vehicle (PAX).

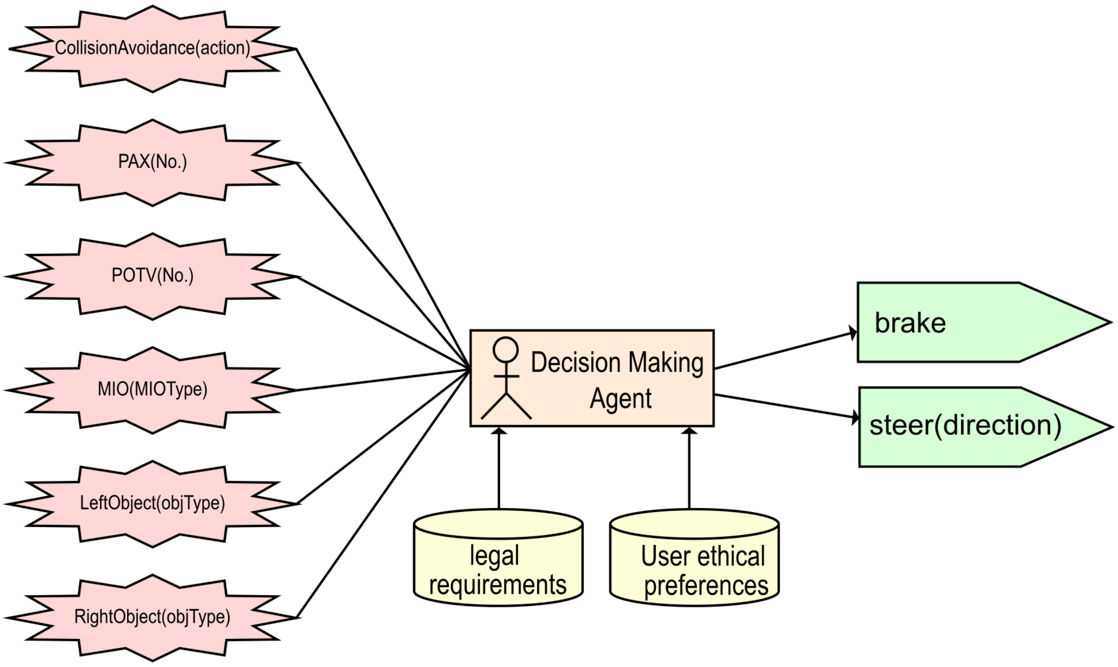

Figure 4.5: Overview of the ethical decision-making system using Prometheus notation.

- `POTV(No.)`: The number of people outside the ego vehicle (POTV).
- `MIO(mioType)`: The most important object, i.e. the one nearest the ego vehicle.
- `LeftObj(objType)`: The type of object to the left of the ego vehicle.
- `RightObj(objType)`: The type of object to the right of the ego vehicle.

Also there are two sources of information. First, the user ethical preferences, and second the legal requirements necessary to operate the system. Finally, we have the two possible actions to implement: braking or steering.

### 4.5.1 Environment perception

In this section, we will explain the ways in which we have implemented the interaction between the environment and the agent in our setup. The class diagram for the environment of the ethical decision-making agent is shown in Figure 4.6. The main class is **WorldEnv**, which inherits from **Environment**, the base environment class provided by Jason. With its methods it initializes the environment, gets instructions from the agent to apply to the vehicle, calls for an update on the environment and checks for the existence of perceptions in the *belief base*. The **VehicleComm** class is used for the communication with the vehicle. This

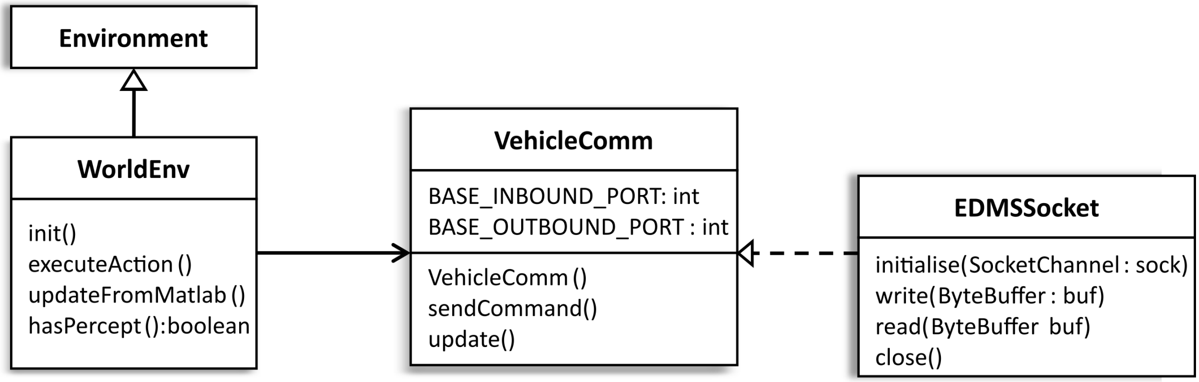

Figure 4.6: Class diagram of the ethical decision-making environment.

implements three methods, one to establish the connexion with the vehicle, and the other two to send and receive data. The sendCommand(int command) method writes the necessary parameters to the buffer to execute the action defined by the agent while the update() method reads the data from the buffer to be used as perceptions. The attributes of this class simply are the number of the port used for the connection with the vehicle. The methods necessary to set up this connection are defined in the **EDMSSocket** class.

Because of the way the reasoning cycle operates (see Figure 4.4), the agent tries to force the perception after each internal action is performed. To overcome this problem we implemented the plans shown in Code fragment 4.1 to force the perception until after an external action, either brake or steer, has been executed. We do this through a set of flags. If the agent does not perceive that a collision avoidance action is necessary, it keeps updating the environment (+update(_): not wait & not done <- updateEnv). When an action is required, a flag is raised to indicate the agent to hold the update until the action has been performed. Once this happens, a belief indicating that the braking (+b_actionImplemented) or steering (+s_actionImplemented) has been implemented is added to the belief base, triggering the environment update (updateEnv).

Code fragment 4.1: Perception update

```
+update(_): not wait & not done <- updateEnv.
+b_actionImplemented(_)<- updateEnv.
+s_actionImplemented(_)<- updateEnv.
```

### 4.5.2 Decision-making

As mentioned before there are two possible actions: brake or steer. As shown in Code fragment 4.2 (Lines 1-3), when the action is to brake `(+collision_avoidance(brake,_))`, the agent sends the instruction to the vehicle control. When the action is to steer (Lines 5-9) `(+collision_avoidance(steer,_))` it has to analyse the scenario and decide the direction; here the agent has three options depending on the ethical preferences and legal requirements: Steer left, steer right or not changing direction.

The latter could appear counter-intuitive as the purpose of the system should be to avoid the collision, however, depending on the selected ethical setting, to continue in the same direction might be the most sensible option. For example, let us say that avoiding all pedestrians by steering is impossible as they are located to the left and right. However, the number of pedestrians that would be harmed is less if the vehicle keeps going forward than if it steers, then it would make sense not to change the direction. Let us keep in mind that avoiding the collision is not possible, yet the agent still has to decide what to collide with and what to avoid.

Code fragment 4.2: Collision avoidance

```
1  +collision_avoidance(brake,_)<-
2                    .print(”Risk of collision detected, braking.”);
3                    brake.
4  
5  +collision_avoidance(steer,_): edms1.mannerOfCollision(Rt,Lft) &
6                     edms1.crashSeverity(Rt, Lft, RtCS, LftCS)
7                     <- +rtCS(RtCS);
8                        +lftCS(LftCS);
9                        !decisionMaking.
```

To decide the steering direction, the agent first defines the manner of collision and the crash severity. This is done through a couple of internal actions such as manner of collision `(mannerOfCollision(Rt, Lft))`. This function returns the manner of collision for the right side `(Rt)` and the left side `(Lft)`. The other one is crash severity `(crashSeverity(Rt, Lft, RtCS, LftCS))`, which takes as inputs the manner of collision and returns the crash severity value for the right `(RtCS)` and left `(LftCS)` sides. These two concepts will be further explained in Sections 4.5.3 and 4.5.4 respectively. The class diagram for them is presented in Figure 4.7. To customise internal actions, we use the **DefaultInternalAction** class provided by the interpreter to simplify the creation of new internal actions. The method DefaultInternalAction is called by Jason to execute the action. The first argument contains information about the current state of the agent and the second argument is necessary in case

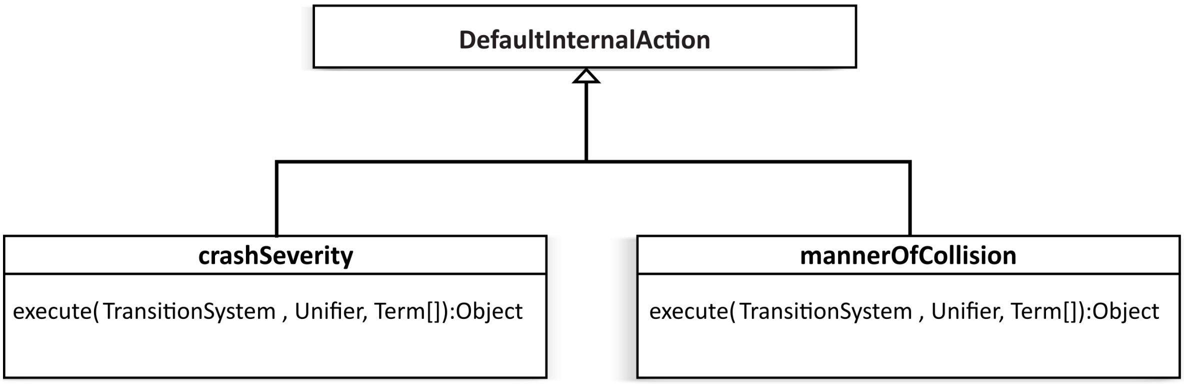
DefaultInternalAction
crashSeverity
execute( TransitionSystem , Unifier, Term[]):Object
mannerOfCollision
execute( TransitionSystem , Unifier, Term[]):Object

Figure 4.7: Internal action class diagram.

the values bounded to any AgentSpeak variables need to be used by the action. The last argument contains an array of terms to be used within the internal action. The output of these functions is used as an input for the ethical decision-making.

When the plan `+!decisionMaking` is added, the agent checks what ethical setting was selected; utilitarian, distributive justice, Kantian, altruism or egoism, and generates a new event to trigger the plan relevant to it as shown in Code fragment 4.3 (Lines 1-6). All the plans handling the selected ethical setting (Lines 8-27) have the right `(RtCS)` and left `(LftCS)` side crash severity result as an input, the plan for the egoist agent `(+!egoistDM)` also considers the number of passengers `(pax)` and the number of people outside the vehicle `(potv)`. The logic of each plan to reach a decision that is induced through an ethical setting is presented in Sections 4.5.6 to 4.5.10. At the end of all the plans, the steering action is executed `<- !steer(D)`, where the `D` represents the direction selected by the agent for the steering.

Code fragment 4.3: decision-making

```

+!decisionMaking <- if(utilitarian_agent) {!utilitarianDM}
                    elif(distributiveJustice) {!distributiveJusticeDM}
                    elif(kantian) {!kantianDM}
                    elif(altruism) {!altruismDM}
                    elif(egoism) {!egoistDM}.

+!utilitarianDM: rtCS(RtCS) & lftCS(LftCS) &
                 edms1.utilitarianDM(RtCS,LftCS,D)
                 <- !steer(D).

+!distributiveJusticeDM: rtCS(RtCS) & lftCS(LftCS) &
                         edms1.distributiveJustice_DM(RtCS,LftCS,D)
                         <- !steer(D).

+!kantianDM: rtCS(RtCS) & lftCS(LftCS) &
                 edms1.kantianDM(RtCS,LftCS,D)
                 <- !steer(D).

+!altruismDM: rtCS(RtCS) & lftCS(LftCS) &
                 edms1.altruistDM(RtCS,LftCS,D)
                 <- !steer(D).

+!egoistDM: rtCS(RtCS) & lftCS(LftCS)
                 & pax(PAX) & potv(POTV) &
                 edms1.egoistDM(RtCS,LftCS,PAX,POTV,D)
                 <- !steer(D).

+!steer(D): true <-
        .print(”Risk of collision detected, steering to the ”,D);
        steer(D).

-!steer(D): true <- .print(” steer action has failed”).
```

### 4.5.3 Manner of Collision

According to the NHTSA [70], the manner of collision is a classification for crashes when the first harmful event involves two vehicles. This describes the way in which two vehicles initially came together. This information is important because, as shown in statistics compiled by the NHTSA on their Fatality Analysis Reporting System (FARS) [71], some types of contact between vehicles or objects, are deadlier than others. In our system, this information is used as a parameter to aid the decision-making. The manners of collision considered in the system are:

- Angle: Two vehicles that impact at an angle. Figure 4.8

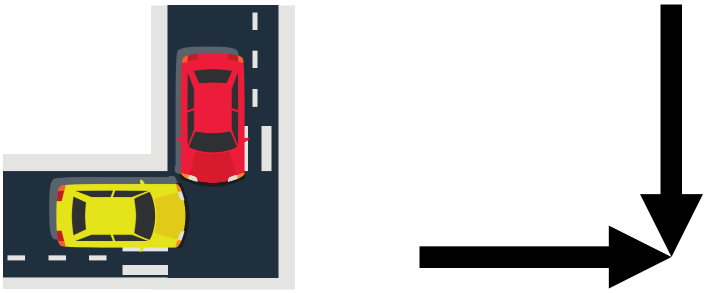

Figure 4.8: Angle Collision.

- Head on: Two vehicles travelling in opposite directions. The front end of one vehicle impacts with the front end of the other. Figure 4.9

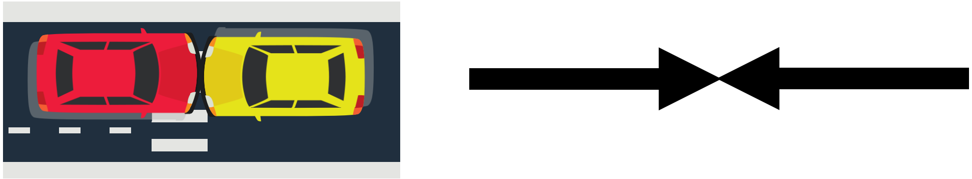

Figure 4.9: Head on collision.

- Rear-end: Two vehicles travelling in the same direction. One vehicle front end collides with the rear end of the other. Figure 4.10

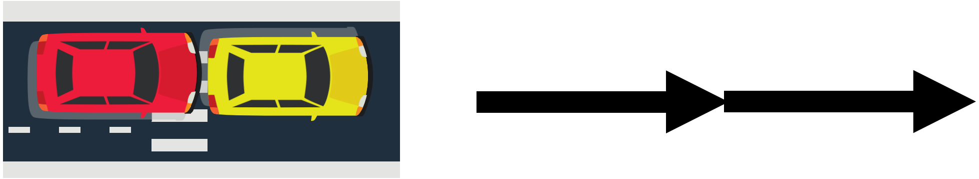

Figure 4.10: Rear-end collision.

- Sideswipe: Two vehicles travelling either in the same or opposite direction. One vehicle impacts the other in a manner that there is no involvement of the front or rear end areas. The impact then swipes along the surface of the other vehicle. Figure 4.11

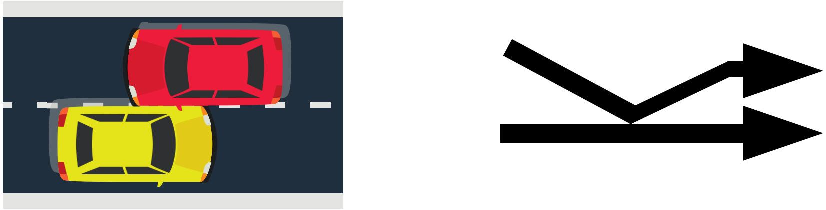

Figure 4.11: Sideswipe collision.

Table 4.1: Crash severity

| **Collision with vehicle** | | **Collision with fixed object** | | **Collision with non-fixed object** | | **Non-collision:** | |
|---|---|---|---|---|---|---|---|
| Angle | 0.17 | Pole/Post | 0.04 | Parked Motor Vehicle | 0.012 | Rollover | 0.078 |
| Rear End | 0.072 | Culvert/Curb/Ditch | 0.067 | Animal | 0.05 | Other/Unknown | 0.012 |
| Sideswipe | 0.027 | Shrubbery/Tree | 0.071 | Pedestrian | 0.17 | | |
| Head On | 0.010 | Other/Unknown | 0.10 | Other/Unknown | 0.039 | | |
| Other/Unknown | 0.05 | | | | | | |

Through the perceptions of the environment, the agent knows if the collision is with a vehicle. If that is the case it determines the type of collision according to both vehicles positions and travel direction.

### 4.5.4 Crash Severity

To determine the crash severity, the system utilizes the manner of collision or the type of the object involved in the accident. Using statistics available in the NHTSA website [71] we are able to assign a numerical value that represents how dangerous a type of crash can be. Higher values mean that particular type of collision has statistically presented more fatalities. The values assigned are normalised in a scale of 0 to 1 and the sum of all values is equal to 1. Such values are presented in Table 4.1.

To determine the crash severity, the system takes into account the manner of collision or the type of object involved in the crash, then it gives a score for the left and right side.

### 4.5.5 Etical Decision-making framework

An overview of the agent in the BDI framework is presented in Figure 4.12. The agent has two beliefs sources. First, there are the internal beliefs of the agent. These refer to the things that the agent knows from the vehicle, like the ethical setting selected by the user or the number of passengers. The second source is the percepts. This is information from the environment, like the number of people outside the vehicle (potv), other vehicles or objects as well as their location. From the percepts, the agent also acquires knowledge about external events that will trigger the decision making process, in this case an event is a collision warning. The desires of the agent are the ethical concerns in the priority order defined for each ethical theory and the legal requirements that it needs to comply with. The plans available to the agent are the actions that it can perform to achieve its goals. In this case, the available actions are to brake, steer to the left or right and keep driving. Once the agent has all this information, it starts the decision-making process. This stage for each ethical theory is explained in the following sections. Once the decision making process selects the applicable plan, this becomes the intention to achieve which should be a manoeuvre that complies with the user setting and the legal requirements. Finally, the agent sends the order to execute the manoeuvre to the vehicle's actuators and updates the belief base.

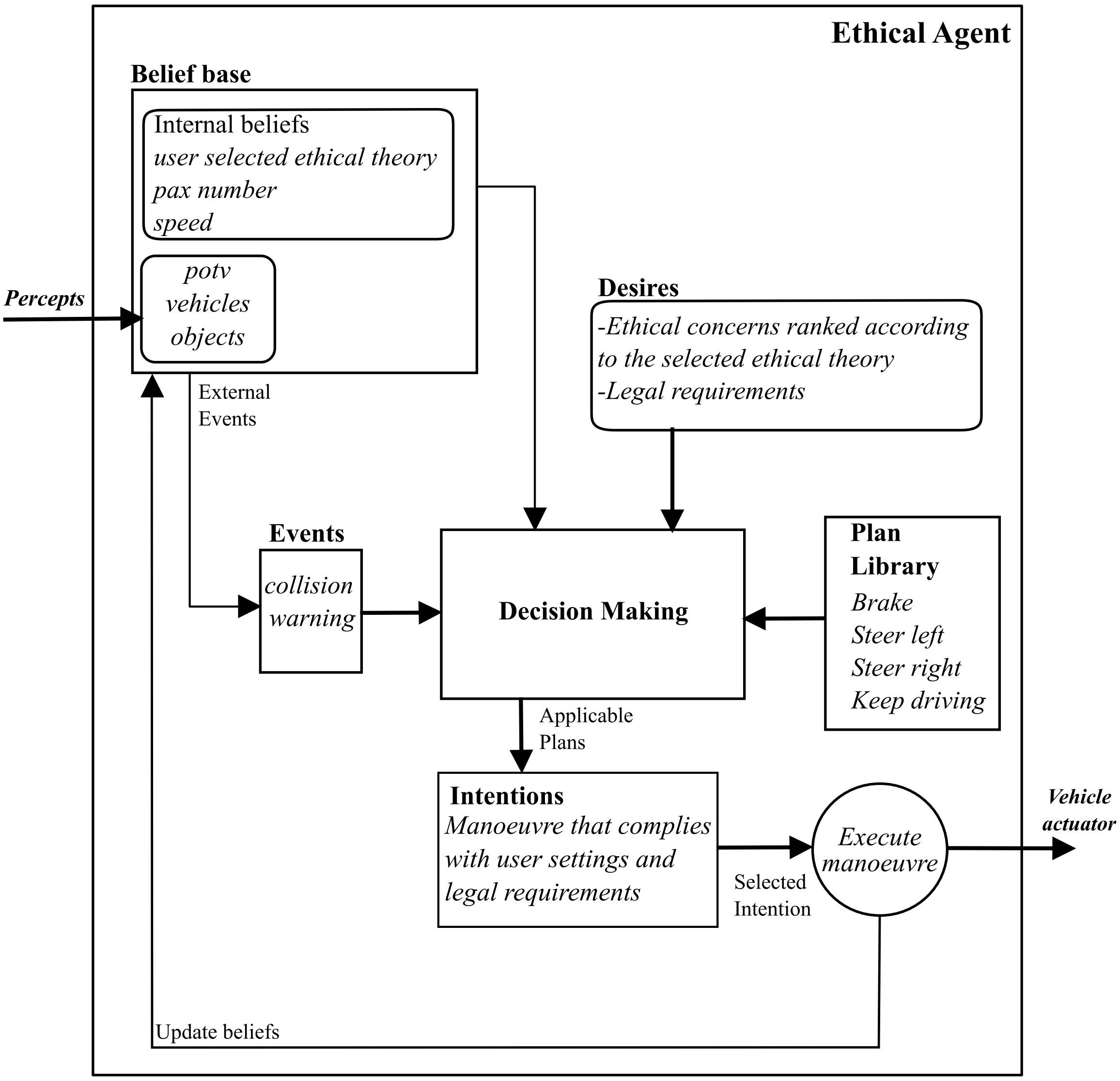

Figure 4.12: Ethical agent framework.

### 4.5.6 Utilitarian decision-making

To derive the final decision for each ethical setting, the agent implements the concerns defined in Chapter 3. From Table 3.1 we have that for the utilitarian option the most important concern is to *harm the least possible number of people*. The logic for this option is presented in Algorithm 1. The algorithm uses as inputs the percepts list $p$ and the crash severity value for left (Left $\alpha$) and right (Right $\alpha$) sides. The first step is to determine which is greater, the number of passengers or the people outside the vehicle, this will help to determine if the priority should be the passengers or the pedestrians. If the priority is the passengers, then it checks if the people outside the vehicle can be safely avoided. If that is the case, then IT proceeds to manoeuvre in that direction. If the priority is the pedestrians, the agent tries to manoeuvre in a manner that the passengers suffer the least damage possible.

To do the above, we implemented a priority score to characterise each type of object. The values assigned are normalised in a scale of 0 to 1 and the sum of all values is equal to 1 as shown in Table 4.2. The higher the priority score is, the less preferred that option is. These values are then added to the *crash severity*, if applicable, resulting in a *severity score* for each direction. For example, if all manoeuvres involve a car crash but one is an angle collision and the other a rear end collision, then the rear end option would be preferable since it is less dangerous. Let us denote the *crash severity* as $\alpha$ and the *priority score* as $\beta$, the *severity score* is calculated as $\gamma = \alpha + \beta$. The preferred manoeuvre is passed to the agent in the form of a direction: left, right or front.

Although the calculation of the severity score $\gamma$ is done as a sum of two quantities $\alpha$ and $\beta$ as proposed above, said calculation could be performed in several ways since it is a problem without an established methodology. For example, another straightforward way to obtain a severity score would be the multiplication of $\alpha$ and $\beta$. We briefly study this approach and show the results of utilizing such calculation in our set up in Chapter 5, Section 5.4.4.

Table 4.2: Priority score values for avoiding objects

| Type of object | Priority score $\beta$ |
|---|---|
| Cannot avoid pedestrians | 0.4 |
| Cannot avoid vehicles | 0.3 |
| Cannot avoid animals | 0.2 |
| Cannot avoid objects | 0.1 |

**Algorithm 1:** Utilitarian decision-making

```
Input  : A set of percepts $p$, Right $\alpha$, Left $\alpha$,
Output: The steering direction $D$
if p.pax $>$ p.potv
    if Free left: True ;
        $D \rightarrow$ Left;
    else if Free right: True ;
        $D \rightarrow$ Right;
    else if Can not avoid pedestrians left and can not avoid pedestrians right ;
        if potvLeft $>$ potvRight ;
        $D \rightarrow$ Do not steer;
    else if Can not avoid pedestrians left and can avoid pedestrians right ;
        $D \rightarrow$ Right;
    else if Can not avoid pedestrians right and can avoid pedestrians left ;
        $D \rightarrow$ Left;
else if p.potv $>$ p.pax
        Calculate $\gamma$ for each $D$;
    if Right $\gamma <$ left $\gamma$ ;
        $D \rightarrow$ Right;
    else $D \rightarrow$ Left;
    end if
end if
```

### 4.5.7 Distributive Justice decision-making

In this setting the most important concern is to *Inflict the least damage possible to people* (Table 3.1). For this, the agent first calculates the *severity score* for each possible direction and then it checks in which direction there would be less people harmed (NoP), irrespectively if they are passengers or pedestrians, and selects the option with the smallest *severity score* and that harms the least possible number of people. The decision process of the distributive justice decision-making is presented in Algorithm 2.

**Algorithm 2:** Distributive justice decision-making algorithm

```
Input  : A set of percepts $p$, Right $\alpha$, Left $\alpha$
Output: The steering direction $D$
if Free left: True ;
    $D \rightarrow$ Left;
else if Free right: True ;
    $D \rightarrow$ Right;
else ;
    Calculate $\gamma$ for each $D$;
    NoP= p.pax+p.potv for each $D$;
    $D \rightarrow$ min $\{NoP_1 * \gamma_1, \dots, NoP_k * \gamma_k\}$ ;
end if;
```

### 4.5.8 Kantian decision-making

For the Kantian setting, the most important concern is *not harm people outside the vehicle* (Table 3.1). First, the agent tries to safely avoid the people outside the vehicle (potv), if this is not possible then steers towards the direction with less pedestrians. If it can avoid all pedestrians but is still colliding with an object, then it calculates the *severity score* and selects the direction with the smallest score. The decision process of the Kantian decision-making is presented in Algorithm 3

**Algorithm 3:** Kantian decision-making algorithm

```
Input  : A set of percepts $p$
Output: The steering direction $D$
if Free left: True ;
    $D \rightarrow$ Left;
else if Free right: True ;
    $D \rightarrow$ Right;
else if Can not avoid pedestrians left and can avoid pedestrians right ;
    $D \rightarrow$ Right;
else if Can not avoid pedestrians right and can avoid pedestrians left ;
    $D \rightarrow$ Left;
else if Can not avoid pedestrians left and can not avoid pedestrians right ;
    $D \rightarrow \min \{p.potv_1, \ldots, p.potv_k\}$;
else if Can avoid pedestrians right and can avoid pedestrians left ;
    Calculate $\gamma$ for each $D$;
    $D \rightarrow \min \{\gamma_1, \ldots, \gamma_k\}$ ;
end if
```

### 4.5.9 Altruism decision-making

For the altruist decision-making the principal concern is *not to harm people outside the vehicle* (Table 3.1). The agent first checks if colliding with potv can be avoided, if not then the objective is to collide with the least possible number of people. If the pedestrians can be avoided, then the agent calculates the *severity score* of each possible direction to select a manoeuvre. The decision process of the altruist decision-making is presented in Algorithm 4.

**Algorithm 4:** Altruist decision-making

```
Input  : A set of percepts $p$, Right $\alpha$, Left $\alpha$
Output: The steering direction $D$
if Free left: True ;
    $D \rightarrow$ Left;
else if Free right: True ;
    $D \rightarrow$ Right;
else if Can not avoid pedestrians left and can avoid pedestrians right ;
    $D \rightarrow$ Right;
else if Can not avoid pedestrians right and can avoid pedestrians left ;
    $D \rightarrow$ Left;
else if Can not avoid pedestrians left and can not avoid pedestrians right ;
    if p.potvLeft >p.potvRight;
        $D \rightarrow$ Right;
    else;
        $D \rightarrow$ Left;
else if Can avoid pedestrians right and can avoid pedestrians left ;
    Calculate $\gamma$ for each $D$;
$D \rightarrow \min\{\gamma_1, \ldots, \gamma_k\}$ ;
end if
```

### 4.5.10 Ethical Egoism decision-making

The egoist setting is very similar to the altruist one, the only difference is on the two first concerns. Here the main concern is to *not harm the passengers* (Table 3.1), however this should be carefully considered, as it might not be legally and morally acceptable to harm a high number of people just to save the passenger. This particular setting can be useful in conjunction with utilitarian setting to solve dilemmas where the number of people harmed is equal on both sides.

In this setting the agent checks if pedestrians can be safely avoided, if not, then it calculates the *severity score* of each direction. If the number of passengers is higher or equal to the number of potv and avoiding them results in a bigger *severity score* than to keep driving in their direction, then the vehicle keeps its trajectory. If the number of pedestrians affected is higher, then the egoist setting can not be applied and the agent instructs the vehicle to perform the manoeuvre that avoids the pedestrians. The decision process of the egoist decision-making is presented in Algorithm 5.

**Algorithm 5:** Ethical Egoism decision-making

```
Input  : A set of percepts $p$, Right $\alpha$, Left $\alpha$
Output: The steering direction $D$
if Free left: True ;
   $D \rightarrow$ Left;
else if Free right: True ;
   $D \rightarrow$ Right;
else Calculate $\gamma$ for each $D$;
end if
if p.pax $\geq$ p.potv ;
      D $\rightarrow$ min $\{\gamma_1, ..., \gamma_k\}$;
   end if;
else if p.potv $\geq$ p.pax ;
   Exclude potv;
      $D \rightarrow$ min $\{\gamma_1, ..., \gamma_k\}$ ;
end if
```

## 4.6 Conclusions

In this chapter a proposal for a system that implements the ethical concerns defined in Chapter 3 has been presented along with a framework for it implementation. This implementation is done through an hybrid architecture with the vehicle control and environment simulation on MATLAB and a rational agent, programmed on AgentSpeak, in charge of the ethical decision-making part. The proposed agent is able to analyse the vehicle surroundings and implement two actions: brake or steer, in case that a collision is detected. To aid this decision, the agent takes into account the ethical concerns of each ethical settings and the *crash severity* of each collision.

The work presented in this chapter carries some challenges and limitations. At the present time, the statistics used to calculate the severity score in Section 4.5.6 do not differentiate between pedestrians and passengers. More research is needed about how to represent the severity score. Performing an accurate representation of the score with the available information is a significant challenge. One of the limitations is that we are assuming that if the car crashes with the pedestrian it would be fatal for them. However, in reality, this is not always the case as it depends on many variables like the speed of the car or the safety systems fitted in the vehicle. Having this information could benefit the system, as one potential consequence of this limitation is that the agent might not get the information needed to get a more accurate decision. The way in which the severity score is presented in this work is just one proposal, but this could be expanded to increase the framework's accuracy.

Another limitation that could arise with the proposed way of calculating the severity score is the fact that we are giving a numerical value to an ethical decision, and that we are doing it while pairing it with our defined concerns. Such quantification might not be enough to justify certain ethical decisions in dilemma situations. Further research is required to derive possible solutions to this limitation such as ways to map this ethical decisions to engineering solutions, as well as more defined legal requirements that can provide a clearer answer for ethical dilemmas.

# Chapter 5

# Simulation and scenario analysis

## 5.1 Introduction

To test the system proposed and defined in Chapter 4, different scenarios were designed. In section 5.3 this scenarios are presented. Section 5.4 shows the outcomes of each ethical theory for each scenario and in section 5.5 a discussion of these results is presented.

## 5.2 Vehicle and Environment Implementation

The vehicle control system and the environment simulation are done in MATLAB. The development of a vehicle control system is outside the scope of this work hence we have used the tools already included within the software to create a simple control system and environment simulation. The MATLAB model is composed by the following modules: vehicle control, vehicle dynamics, actors and sensor simulation, tracking and sensor fusion and communication, as shown in Figure 5.1. The communication module is the one that interacts with the ethical agent presented in Chapter 4. These modules are explained in more detail in this section.

The different scenarios for the simulation, further explained in this chapter, were created using the *Driving Scenario Designer* application in MATLAB. This allows us to define the road, the *ego vehicle* which is the vehicle with the decision-making agent, and the *actors*, which encompass all the other participants in a scenario. The possible actors are divided in 5 different types: car, truck, bicycle, pedestrian and barrier. Vision and radar sensors characteristics and location on the vehicle can also be defined in the application.

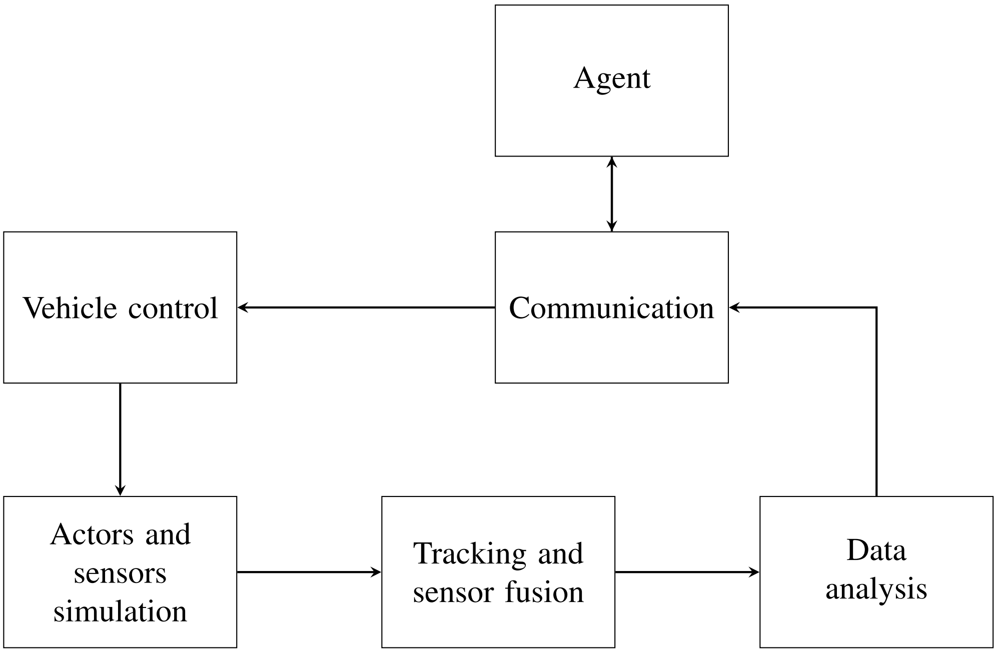

Figure 5.1: System Matlab model

In our simulations we can identify two coordinate systems: The world coordinate system, a universal system where all the actors are placed, and the vehicle coordinate system, which is anchored to the vehicle. In this case the centre is located in the middle of the rear axis.

The position of the ego vehicle and all actors can be expressed in relation to the ego vehicle, we call this the vehicle coordinate system.

**Vehicle Control System** The control system comprises the vehicle control and vehicle dynamics modules. For the vehicle dynamics module we utilise the Simulink block *vehicle body 3DOF single track*. This module allow us to calculate longitudinal, lateral, and yaw motion and is based on the bicycle model. The bicycle model simplifies the four wheels of

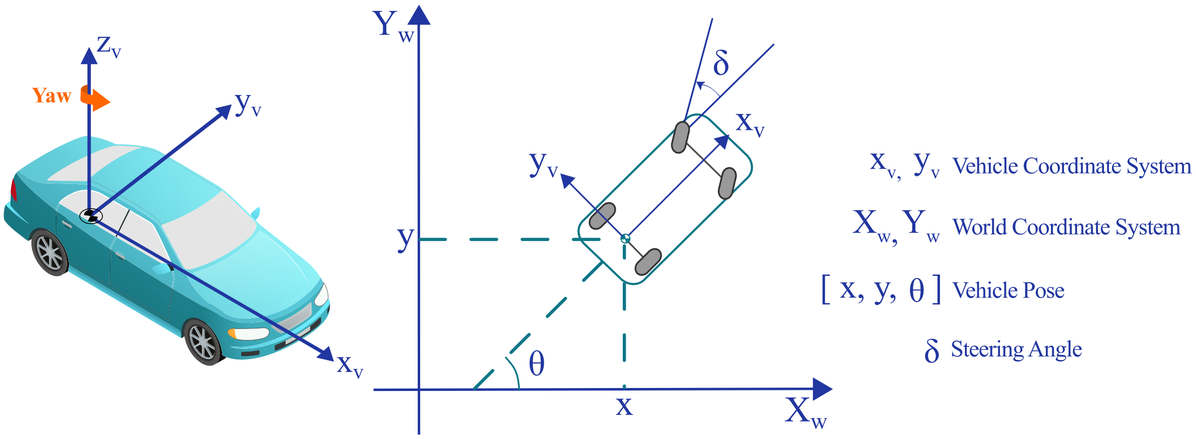
$z_v$
Yaw
$y_v$
$x_v$
$Y_w$
$\delta$
$x_v$
$y_v$
y
$\theta$
x
$X_w$
$x_v, y_v$ Vehicle Coordinate System
$X_w, Y_w$ World Coordinate System
$[x, y, \theta]$ Vehicle Pose
$\delta$ Steering Angle

Figure 5.2: Coordinate systems for the simulation.

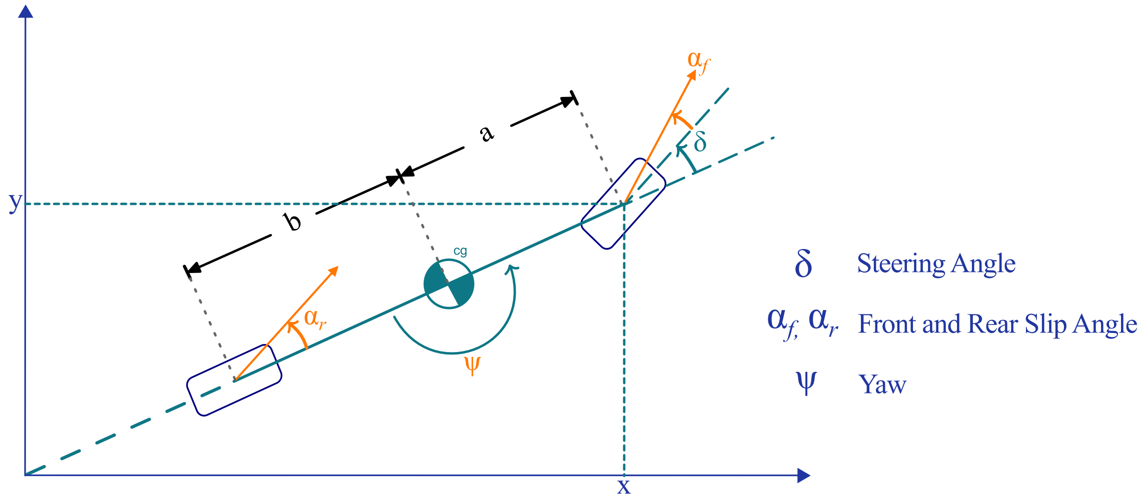

Figure 5.3: 3-DOF Bicycle model used for simulation.

the vehicle into two wheels, combining the front and rear wheels respectively, as shown in Fig. 5.3. This model is widely used for the modelling of vehicle systems, the equations of motion for this model are readily available in the literature [72, 73]. In the following sections we include the ones used by the Simulink block.

The system has three inputs; the front wheel steering angle, and the front and rear forces. These allow us to control the steering of the vehicle as well as the acceleration and brake operations. As outputs, we obtain the $x$ and $y$ displacement of the wheel along the $x$ and $y$ earth-fixed axes, the lateral and angular velocities $\dot{y}$ and $\dot{\psi}$ and the yaw angle $\psi$. These are measured in m, m/s, rad and m/s$^2$ respectively.

To calculate the dynamics of the vehicle, the block uses the following equations:

$$\ddot{y} = -\dot{x}r + \frac{F_{yf} + F_{yr}}{m}, \tag{5.1}$$

$$\dot{r} = \frac{aF_{yf} - bF_{yr}}{I_{zz}}, \tag{5.2}$$

$$r = \dot{\psi}, \tag{5.3}$$

$$\ddot{x} = \dot{y}r + \frac{F_{xfinput} + F_{xrinput}}{m}, \tag{5.4}$$

where $\ddot{y}$ is the lateral acceleration, $\dot{x}$ is the longitudinal velocity, $F_{yf}$ and $F_{yr}$ are the lateral tire forces applied to front and rear wheels respectively, $m$ is the vehicle mass, $a$ and $b$ are the distances from the front and rear wheels to the vehicle centre of gravity CG, $r$ is the angular velocity, $I_{zz}$ is the moment of inertia of the vehicle body, $\ddot{x}$ is the longitudinal acceleration, $\dot{y}$ is the lateral velocity and $F_{xfinput}$ and $F_{xrinput}$ are the longitudinal forces applied to the front and rear wheels. These last two are provided by the throttle and brake control that communicates with the ethical agent to control the vehicle.

To calculate the slip angles of the front $\alpha_f$ and rear $\alpha_r$ wheels, the block uses the ratio of the local and longitudinal and lateral velocities $\dot{x}$ and $\dot{y}$ and the front and rear steering angles, $\delta_f$ and $\delta_r$:

$$\alpha_f = atan\left(\frac{\dot{y}+ar}{\dot{x}}\right) - \delta_f, \tag{5.5}$$

$$\alpha_r = atan\left(\frac{\dot{y}+br}{\dot{x}}\right) - \delta_r. \tag{5.6}$$

The tire forces $F_{yf}$ and $F_{yr}$ are

$$F_{yf} = -C_{yf}\alpha_f\mu_f\frac{F_{zf}}{F_{znom}}, \tag{5.7}$$

$$F_{yr} = -C_{yr}\alpha_r\mu_r\frac{F_{zr}}{F_{znom}}, \tag{5.8}$$

where $C_{yf}$ and $C_{yr}$ are the front and rear wheel cornering stiffness and $\mu_f$ and $\mu_r$ are the front and rear wheel friction coefficient.

To maintain the pitch and roll equilibrium, the block calculates the normal force applied to front $F_{zf}$ and rear $F_{zf}$ wheels as follows:

$$F_{zf} = \frac{bmg-(\ddot{x}-\dot{y}r)mh}{a+b}, \tag{5.9}$$

$$F_{zr} = \frac{amg-(\ddot{x}-\dot{y}r)mh}{a+b}, \tag{5.10}$$

where $h$ is the height of the vehicle CG.

The vehicle parameters used for simulation are presented in Table 5.1. The values of the friction coefficient $\mu_f$, $\mu_r$ and the cornering stiffness $C_{yf}, C_{yr}$ used for this simulation are the ones provided as default by Matlab.

**Actors and sensors simulation** The outputs from the vehicle control module are passed

Table 5.1: Vehicle parameters used for simulation.

| Parameter | Value |
|---|---|
| Mass $m$ | 1575 kg |
| Inertia (about Z-axis) $I_{zz}$ | 2875 $kgm^2$ |
| Front Axle to centre of gravity $a$ | 1.2 m |
| Rear Axle to centre of gravity $b$ | 1.6 m |
| Height of vehicle centre of gravity $h$ | 0.35 m |
| Friction coefficient $\mu_f$, $\mu_r$ | 1 |
| Front tire corner stiffness $C_{yf}$ | 12000 N/rad |
| Rear tire corner stiffness $C_{yr}$ | 11000 N/rad |

to the *Scenario Reader* block. This reads the ego vehicle data and a *Driving Scenario* object which contains information about the road, such as its width and number of lanes, and the actors i.e. number of actors, physical characteristics (width, length) and type of object. Then this information is passed to the *Vision Detection Generator* and *Radar Detection Generator* blocks, these blocks allow the system to generate vision and radar detections from the cameras and radar sensors mounted on the ego vehicle. These detections are obtained from the simulated actor poses and are referenced to the ego vehicle coordinate system.

To simplify our model, the vehicle only uses one vision and one radar sensor. For the sensing, the vehicle employs a vision sensor and a radar sensor. Both of these are the standard implementation available on the Automated Driving Toolbox in Simulink with the default values. Both sensors x and y position are located in coordinates (0,0) with respect to the ego vehicle coordinate system shown in Fig 5.2.

**Tracking and sensor fusion** The information generated in the previous step is used as the input of the *Multi-Object Tracker block*. It assigns the detections to tracks which at the beginning are considered temporal. Once enough detections have been assigned to the same track, this is considered to be a physical object and changes it status to confirmed. From confirmed tracks we can also obtain its object type.

Once all the tracks are in place, this information is used to determine the Most Important Object (MIO). This is defined as the one that is located closest to the vehicle and within the same lane.

**Data Analysis** In this module, all the information that will be sent to the agent is compiled and prepared. Based on the position of each actor, the system determines their position in reference to the vehicle. To reduce the amount of information sent to the agent, we stablish that the agent only needs knowledge about the ego vehicle position and speed, and the nearest actor to the left, front and right. For these actors, information about their position, classification (the type of actor), travel direction and if it is moving or not it is also send to the agent.

**Communication Module** This module oversees the bidirectional communication with the agent. It is based on the communication system presented in [64, 74]. The communication is achieved through a TCP connection using sockets in Matlab and Java. The transmission is done using the TCPSend block and the reception with a TCPReceive block available in Simulink. The flow of data is shown in Fig 5.4.

On the Matlab side, we created a buffer with all the information prepared in the data analysis module. This data is sent to the abstraction layer in Java and assigned to a type. We identify two types of data: ego vehicle data (i.e. number of passengers) and perception data (i.e. position and actor type). This information is later used by the agent as the source of its perceptions. On the other side, Java communicates the agent decisions to the vehicle control in Matlab. The information sent is a command number that is interpreted by Matlab as 1 to steer to the left, 2 steer to the right and 3 to brake. This is all the information needed by the vehicle control which will determine the steering angle and break force based on the vehicle dynamics.

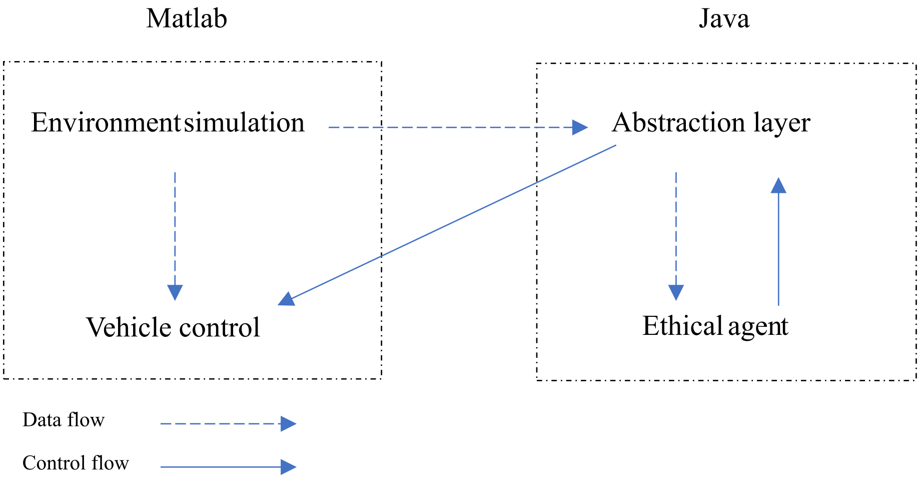

Figure 5.4: System communication diagram.

## 5.3 Scenarios Definition

Three different scenarios were defined to test the system, as shown in Figure 5.5, all scenarios are defined in a four lane road, each lane has a width of 3.2 m. In the first scenario, labelled I, the ego vehicle travels in the second lane from the right, a pedestrian is crossing the road from right to left and there is a second vehicle to the right of the ego. In this scenario the left lane is empty.

The second scenario, labelled II, is similar to scenario I with the lorry travelling in the opposite direction to the ego vehicle at a constant velocity, and it is assumed that it will not change its trajectory. In this scenario, there are two pedestrians to the right of the vehicle, only one crosses the road while the other stays in the right-hand lane.

The third scenario, labelled III, is similar to the first with the ego vehicle, a second vehicle to the right of the ego and a pedestrian in the same position. On the two lanes to the right the vehicles travel in the same direction while the two lanes to the left travel in opposite direction. A lorry on the first lane from the right travels in opposite direction to the ego vehicle. A pedestrian suddenly appears between the cars in the right lane and crosses the road. This third scenario has two different setups, let us call them setting A, where there is one passenger and two pedestrians crossing the road, and setting B, with two passengers travelling in the vehicle and one pedestrian crossing the road. These two settings allows us to evaluate the system when there are more passengers than pedestrians.

The physical characteristics of the ego vehicle used for the simulation are as follows: 4.7 m length, 1.8 m width and 1.4 m height. These measurements correspond to a saloon vehicle similar to an Audi A4 [75]. The initial conditions for the ego vehicle for the different scenarios are presented in Table 5.2.

The other actors in the scenarios are two vehicles of similar size to the ego vehicle, and a lorry of 8.2 m length, 2.5 m width and 3.5 m heigh that maintains a constant speed of 4 m/s. The pedestrians have a profile of 1.7m in height, 0.6 m length and 0.5 m width, and move at a constant speed of 1.5 m/s. The spatial location of the actors can be defined in reference to the vehicle coordinate system, as shown in Table 5.3, where the $x, y$ coordinates correspond to the centre of the rear axe of each vehicle. A graphic representation of these initial conditions is shown in Fig 5.6.

Having defined our scenarios, we are now ready to implement our simulations. These are shown in the following section.

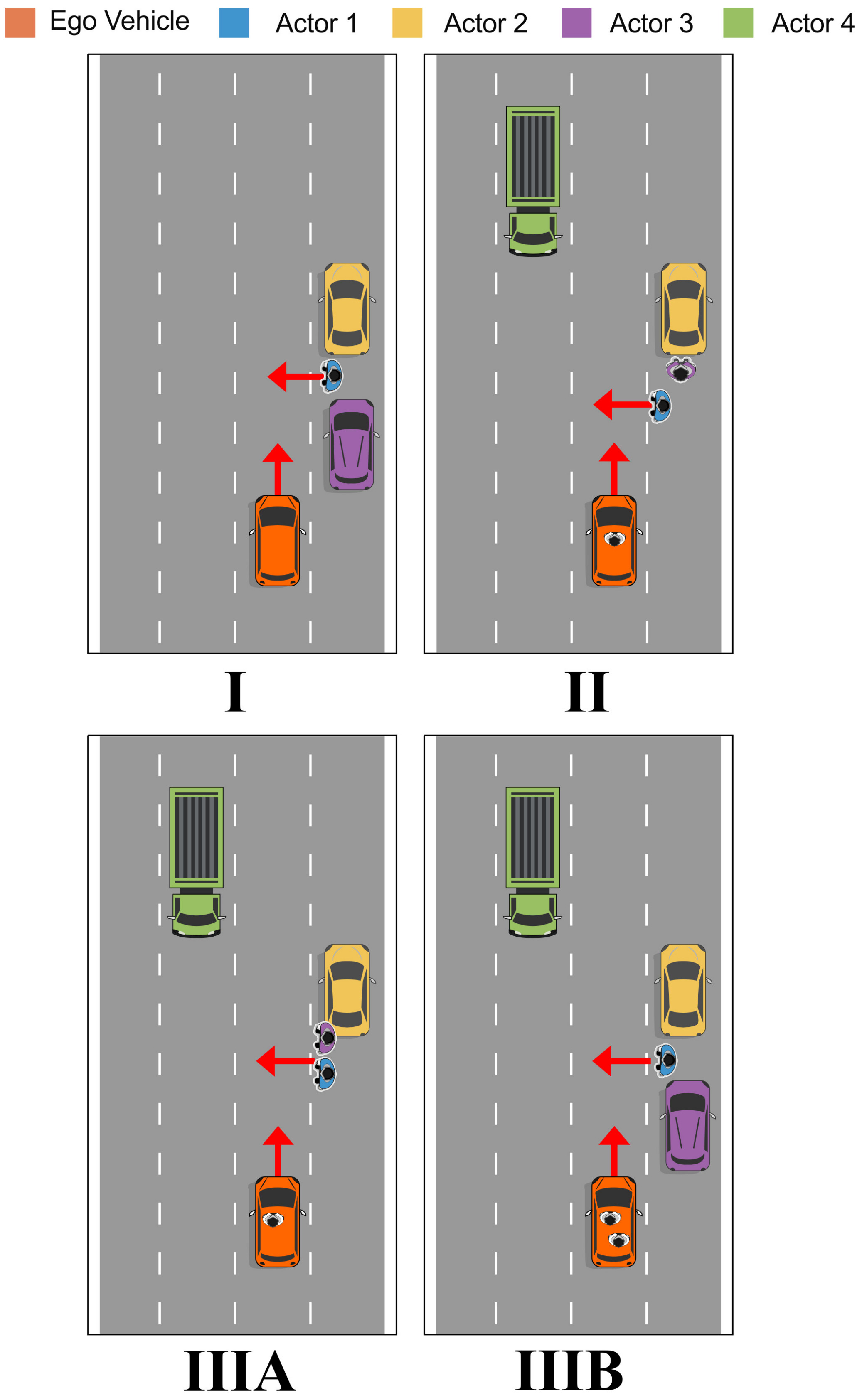

Figure 5.5: Scenarios to test the system. The orange car represents the ego vehicle, the red arrows represent the direction in which the actors are moving.

Table 5.2: Ego vehicle initial conditions for all scenarios.

| | Scenario I | Scenario II | Scenario IIIA | Scenario IIIB |
|---|---|---|---|---|
| Number of passengers | 2 | 1 | 1 | 2 |
| Ego vehicle initial speed (m/s) | 14 | 14 | 14 | 14 |
| Ego vehicle Initial yaw angle (rad) | 0 | 0 | 0 | 0 |

Table 5.3: Actors initial conditions for all scenarios.

| | Actor 1 | Actor 2 | Actor 3 | Actor 4 |
|---|---|---|---|---|
| Type of actor | Pedestrian | Car | Car | Lorry |
| Initial x position (m) | 18.36 | 20.55 | 7.1 | 47.7 |
| Initial y position (m) | -5.62 | -3.09 | -3.24 | 3.41 |
| Initial yaw angle (rad) | 0 | 0 | 0 | 0 |

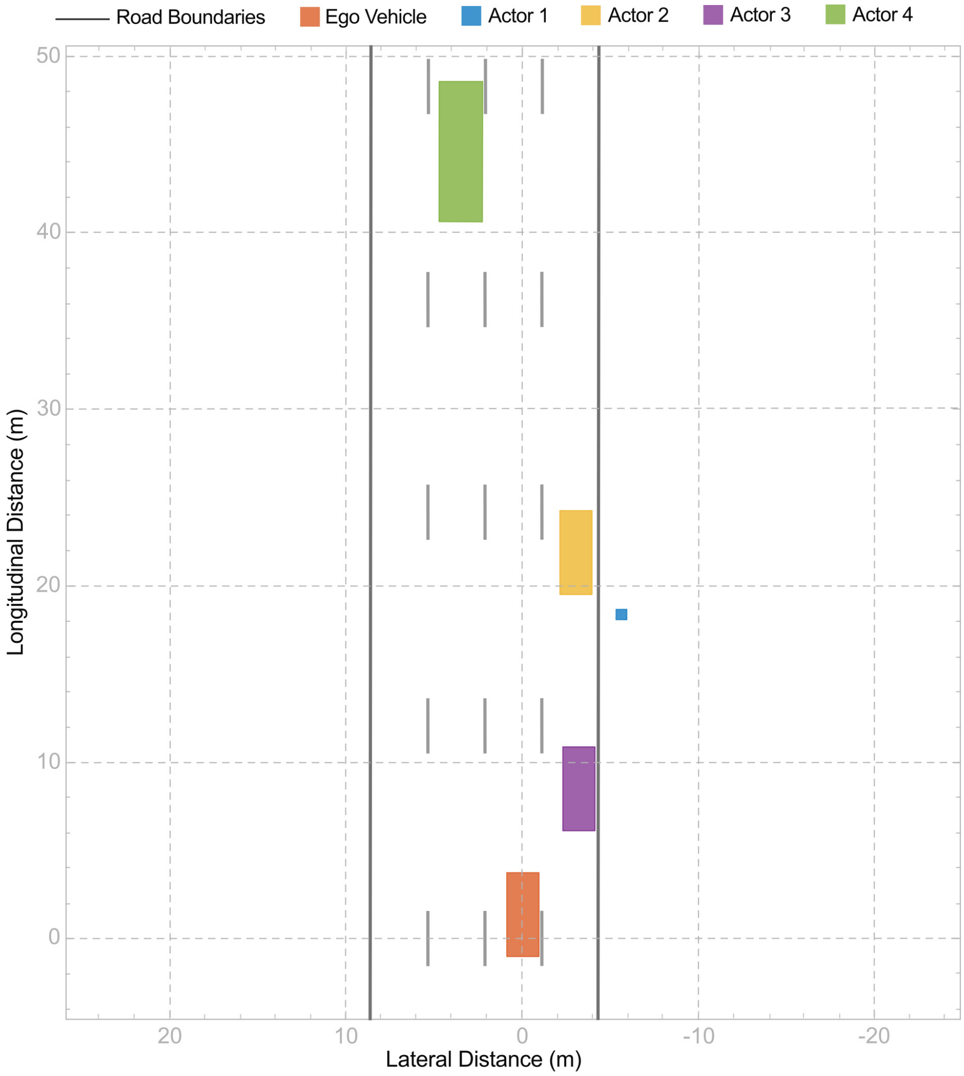

Figure 5.6: Initial physical conditions for all actors

## 5.4 Results

In this section, the results for each scenario are presented. Each scenario was run on all the ethical modes to compare the ways in which each of them react to the same conditions.

It is worth to mention that for all the scenarios presented, the agent always detects that the first response action is to brake. As shown in the messages recorded in the agent console 5.1. The agent uses the time to collision (TTC) as a parameter, it is when there is not enough time to brake that it decides to steer.

Agent console 5.1: Utilitarian agent, scenario IA

```
TTC: -0.8
[ethicalAgent] Risk of collision detected, braking.
```

### 5.4.1 Scenario I

As presented in Section 5.3, the first scenario consist of the ego vehicle travelling northbound in its lane when a pedestrian suddenly crosses the road. Additional, there is also a parked vehicle to the right of the ego and the left lanes are free.

This scenario proves that the agent will always select the safest manoeuvre. The result is the same for all the ethical settings and is as shown in Fig. 5.7. Since the results are similar and to avoid repetition, the severity score obtained in each ethical setting is summarised in Table 5.4.

Table 5.4: Severity scorer results of scenario I for all agents

| | Right severity score | Left severity score |
|---|---|---|
| Utilitarian agent | 0.012 | 0 |
| Distributive Justice agent | 2.327 | 0 |
| Kantian agent | 0.327 | 0 |
| Altruist agent | 0.327 | 0 |
| Ethical egoist agent | 0.037 | 0 |

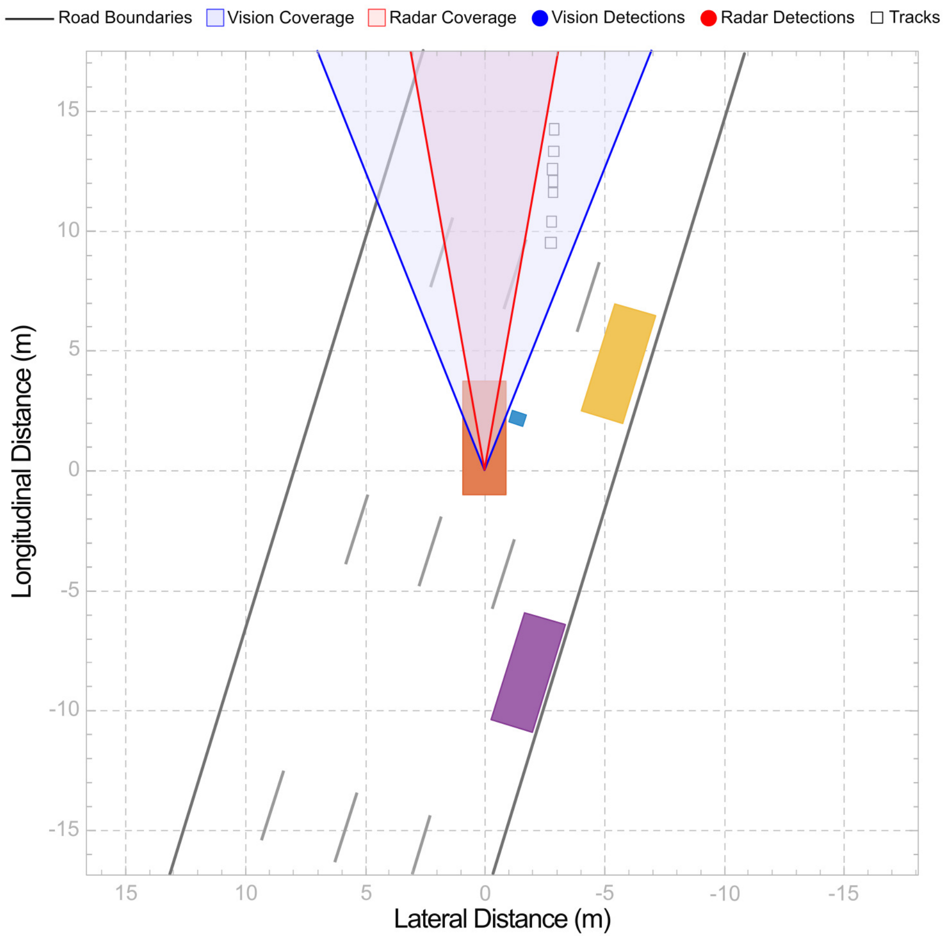

Figure 5.7: Utilitarian theory result scenario I. Turning left is the option that has the lowest severity score, hence the ego vehicle turns left. The results obtained with all the others ethical settings is the same, always turns to the left.

## 5.4.2 Scenario II

### Agent with utilitarian beliefs

In scenario II, there are two pedestrians in different locations, while in the vehicle there are two passengers. The agent determines that the collision to the left is a head on collision (`lftHO`) and the priority is to protect the passengers (`protect PAX`). Then it calculates the severity score for the left (`lftScore`) and right (`rtScore`) side as shown in the Agent console 5.2. Since it cannot avoid the pedestrians without harming both passengers, the agent decides not to change direction, harming only the pedestrian that appeared suddenly in its way, as illustrated in Fig. 5.8.

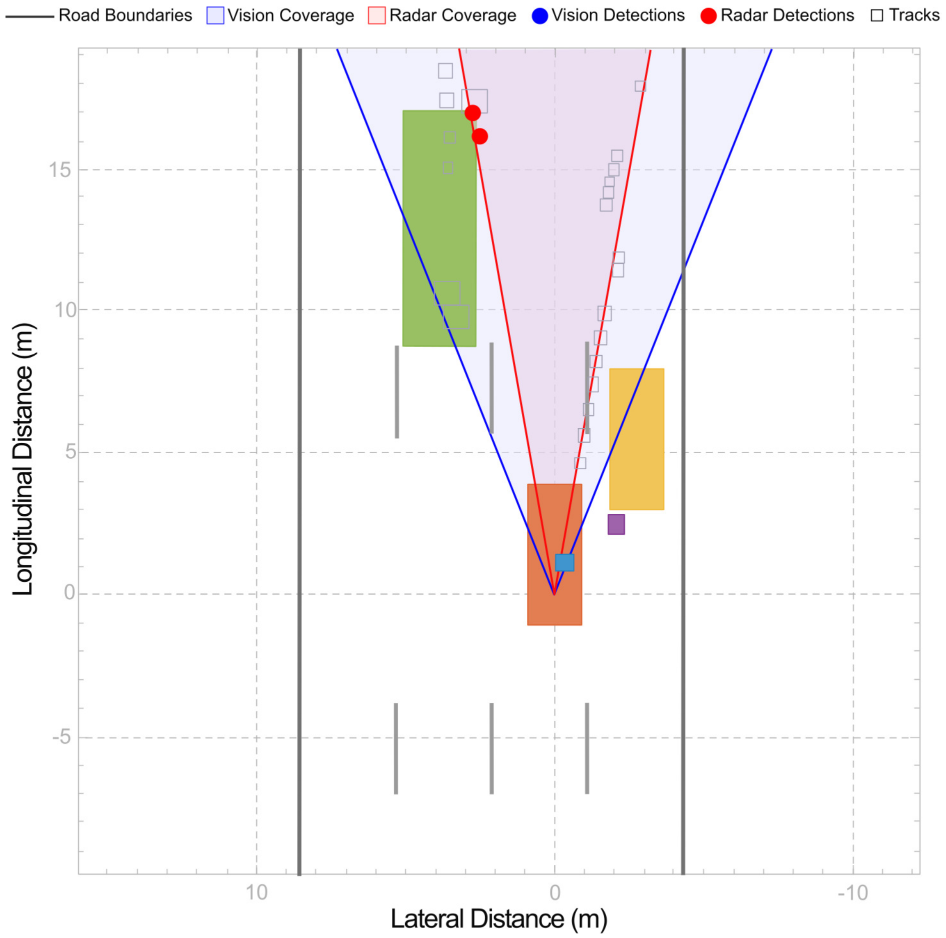

Figure 5.8: Utilitarian theory result scenario II. In this scenario the pedestrians (in black and green) can not be avoided without harming the passengers and since they are the priority the vehicle decides not to change the direction, harming only one person.

Agent console 5.2: Utilitarian agent, scenario II

```
 --manner of collision--
 lftHO
**utilitarian DM**
Priority: protect PAX
rtScore: 0.35 lftScore: 0.3
Severity score
RtCS: 0.35, LFftCS: 0.372
[ethicalAgent] Risk of collision detected, keep driving.
```

### Agent with distributive justice beliefs

In this ethical theory, the agent determines how much people would be involved in each possible manoeuvre, determines the manner of collision (`lftHO`) and calculates the severity score for each for the right (`RtCS`) and left (`LFftCS`) sides, as shown in the Agent console 5.3. For this case, the option that involves the smallest severity score and the least number of people is turning to the left, illustrated in Fig. 5.9.

Agent console 5.3: Distributive justice agent, scenario II

```
 --manner of collision--
 lftHO
**Distributive justice DM**
 People affected if turn to the left: 2
 People affected if turn to the right: 3
 People affected if not change direction: 3
 Severity score
 RtCS: 3.35, LFftCS: 2.372
[ethicalAgent] Risk of collision detected, steering to the left
```

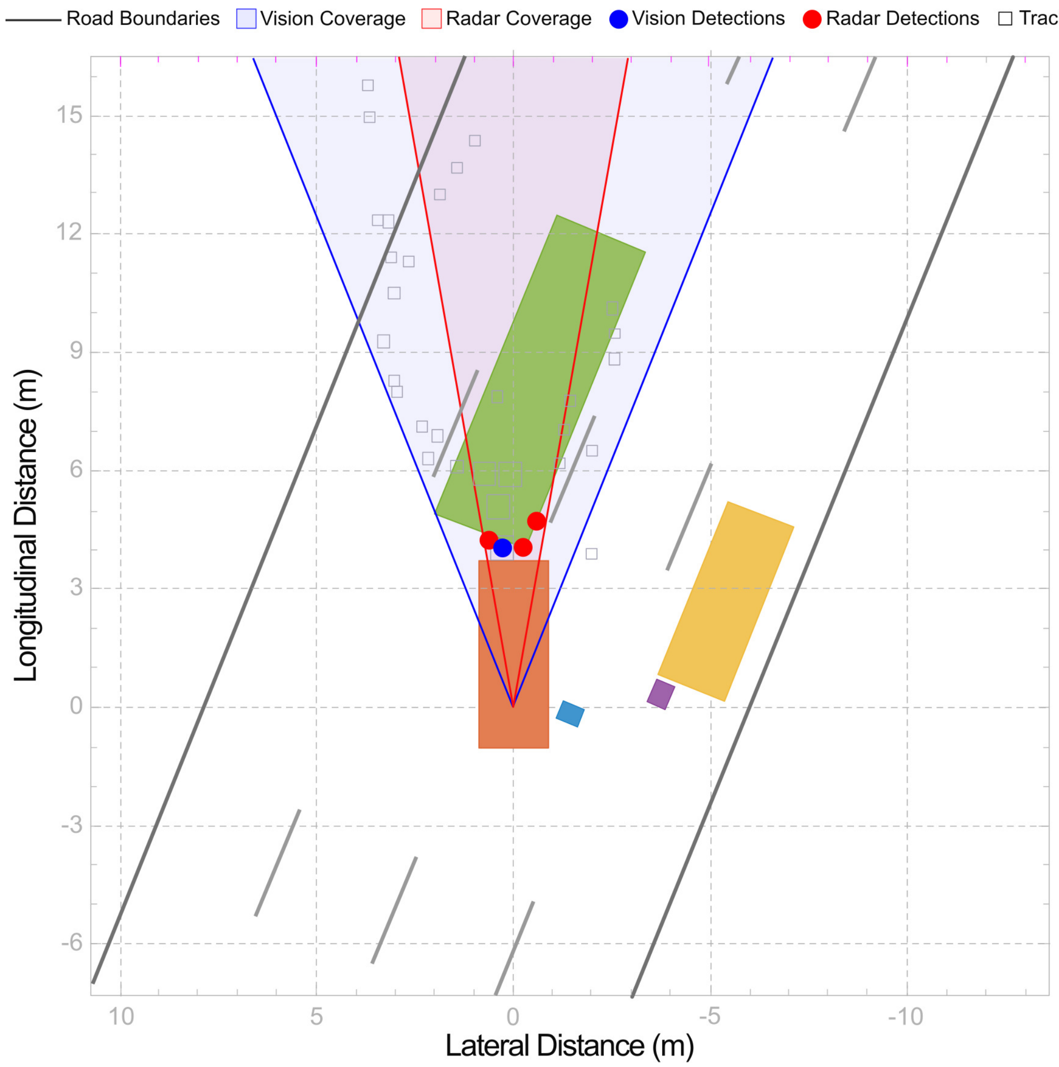

Figure 5.9: Distributive justice theory result scenario II.

### Agent with Kantian beliefs

In this scenario, the Kantian agent behaves as expected from Table 3.1, avoiding to involve the people outside the vehicle. As shown in the Agent console 5.4, the agent first determines the manner of collision for the right side as a rear-end collision (`rtRE`) and a head on collision for the left side (`lftHO`). It also determines that by turning to the left it can avoid the pedestrian, hence this action is implemented, as illustrated in Fig. 5.10.

Agent console 5.4: Kantian agent, scenario II

```
 --manner of collision--
 rtRE
 lftHO
**Kantian DM**
Can avoid pedestrians left
Can not avoid pedestrian right
Can not avoid pedestrians front
[ethicalAgent] Risk of collision detected, steering to the left
```

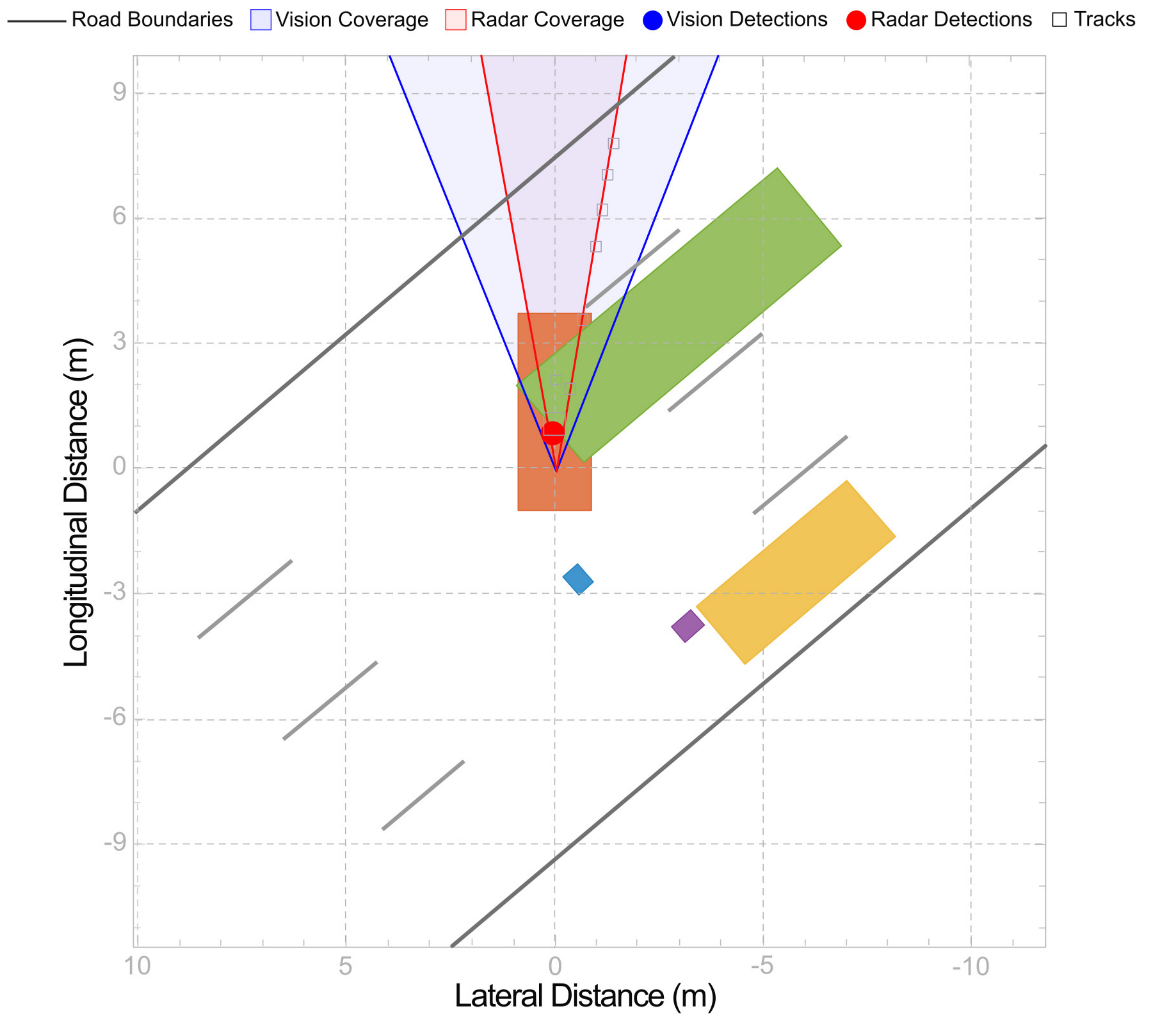

Figure 5.10: Kantian theory result scenario II. In this situation, turning left is the only option where the vehicle can avoid the pedestrians, hence this is the selected option.

### Agent with altruist beliefs

In the case of the altruist agent, the priority in this scenario is to protect the pedestrians. First, it determines that to the left there is a head-on collision (`lftHO`) and to the right a collision with a parked vehicle (`rtPV`). Also the agent determines that the only option to avoid the pedestrians is to turn to the left, as shown in the Agent console 5.5, so it applies this action even though turning left is more dangerous to the passengers as illustrated in figure 5.11.

Agent console 5.5: Altruist agent, scenario II

```
 --manner of collision--
 lftHO
 rtPV
**altruist DM**
Can avoid pedestrians left
Can  not avoid pedestrian right
Can not avoid pedestrians front
 Severity score
 RtCS: 0.35, LFftCS: 0.372
[ethicalAgent] Risk of collision detected, steering to the left
```

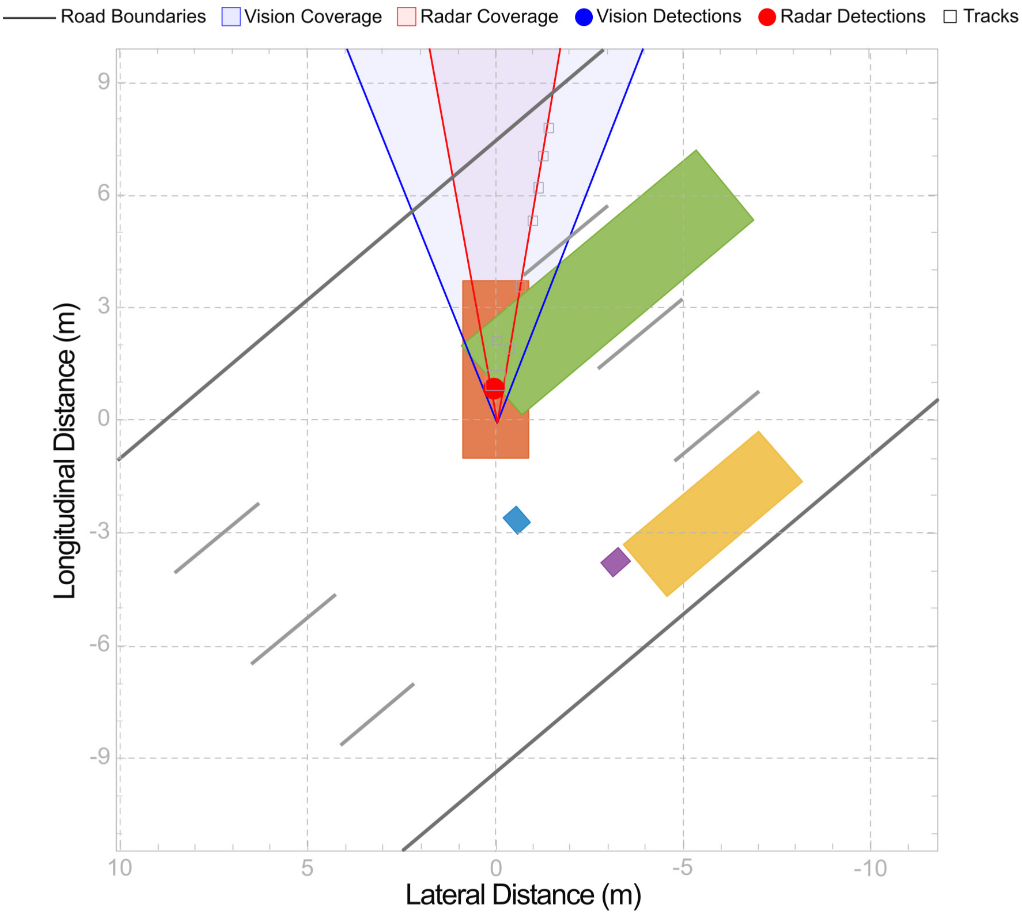

Figure 5.11: Altruist theory result scenario II. Even though turning to the left is more dangerous for the passengers, the altruist agent decides to protect the pedestrians.

### Agent with ethical egoism beliefs

In this scenario, for the ethical egoism theory, the agent determines that the priority are the passengers, because they are more in number than the pedestrians affected, as show in the Agent console 5.6. To the left there is a head-on collision (`lftHO`) and to the right a collision with a parked vehicle (`rtPV`). Turning left has a higher severity score, by turning right it would still collide with a pedestrian. So, the agent decides not to change direction as illustrated in Fig. 5.12

Agent console 5.6: Egoist agent, scenario II

```
 --manner of collision--
 lftHO
 rtPV
**egoist DM**
  Priority: protect PAX
  Severity score
 RtCS: 0.35, LFftCS: 0.372
[ethicalAgent] Risk of collision detected, keep driving
```

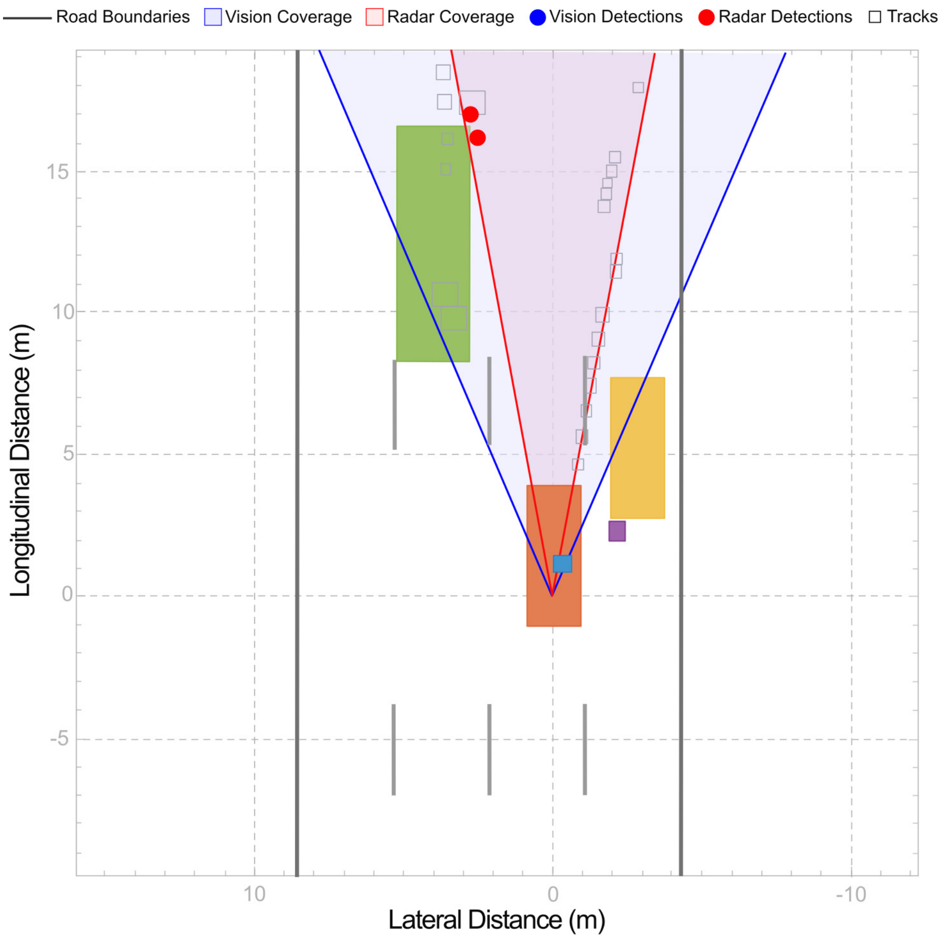

Figure 5.12: Egoist theory result scenario II.

### 5.4.3 Scenarios IIIA and IIIB

#### Agent with utilitarian beliefs

For the utilitarian theory on the scenario III setting A, as explained in Section 3.2, Table 3.1, the priority is to protect the pedestrians as they are greater in number than the passengers. The agent calculates that the collision to the left is a head on collision (`lftHO` ) with a severity score of 0.322. To the right it results in a rear end collision lftHO with a severity score of 0.277. The severity score to the right is lower, so it turns the vehicle to the right (Fig. 5.13). The results from the agent processing are shown on the agent console 5.7

Agent console 5.7: Utilitarian agent, scenario IIIA

```
--manner of collision--
 rtRE
 lftHO
**utilitarian DM**
  Severity score
 RtCS: 0.277, LFftCS: 0.322
[ethicalAgent] Risk of collision detected, steering to the right
```

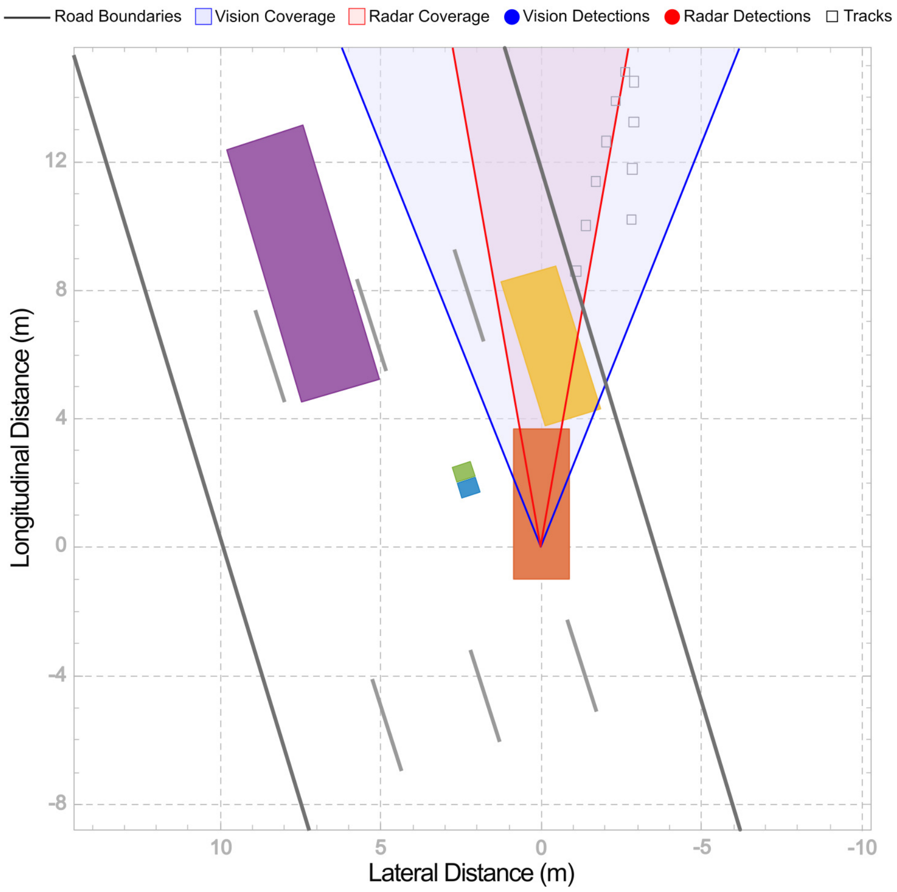

Figure 5.13: Utilitarian theory result scenario IIIA. The vehicle priority is the pedestrians (green and black), from the two options available the less risky is to turn to the right.

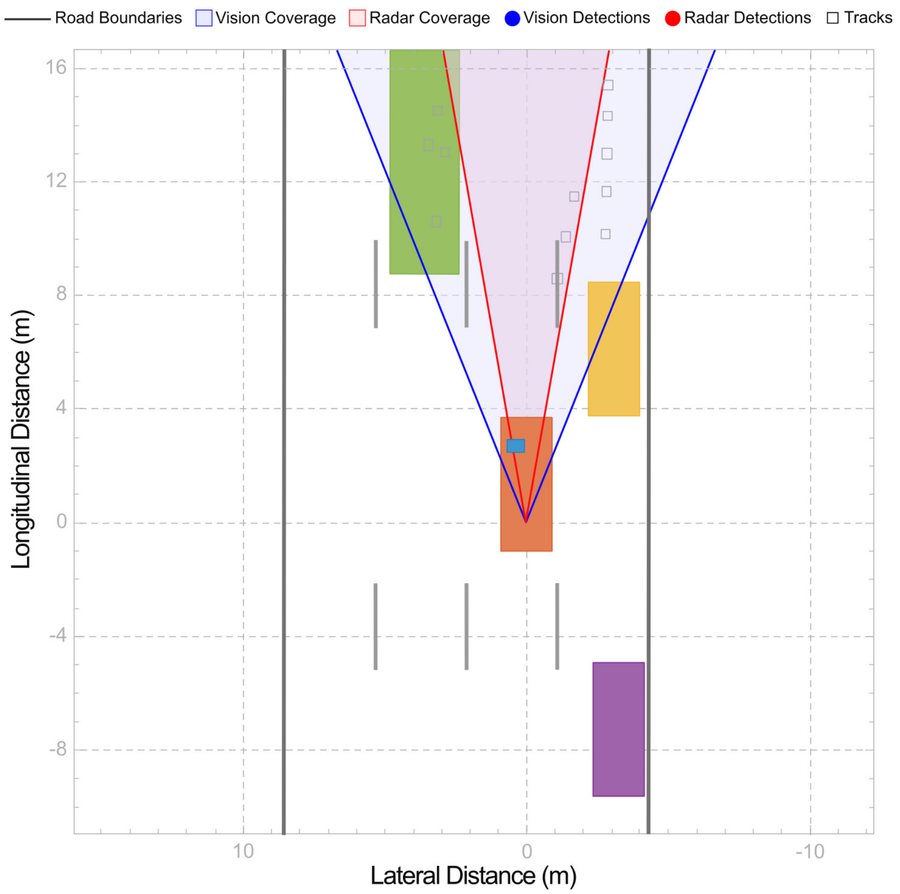

Figure 5.14: Utilitarian theory result for scenario IIIB. The vehicle priority is the passengers, reason why it decides not to change of direction, harming the pedestrians instead of the passengers.

For the scenario IIIB, the agent determines that the priority is to protect the passengers, as now there are more in number than the pedestrians. In this case it is not possible to avoid the pedestrian without harming the passengers. The agent decides to keep driving without changing direction (Fig. 5.14) as shown in the agent console 5.8.

Agent console 5.8: Utilitarian agent, scenario IIIB

```
 --manner of collision--
 rtPV
 lftHO
**utilitarian DM**
  Priority: protect PAX
  Severity score
 RtCS: 0.312, LFftCS: 0.372
[ethicalAgent] Risk of collision detected, keep driving
```

From the above, the decision-making process coincides with the priorities introduced in Section 3.2 i.e. the actual outcome is equal to the expected outcome.

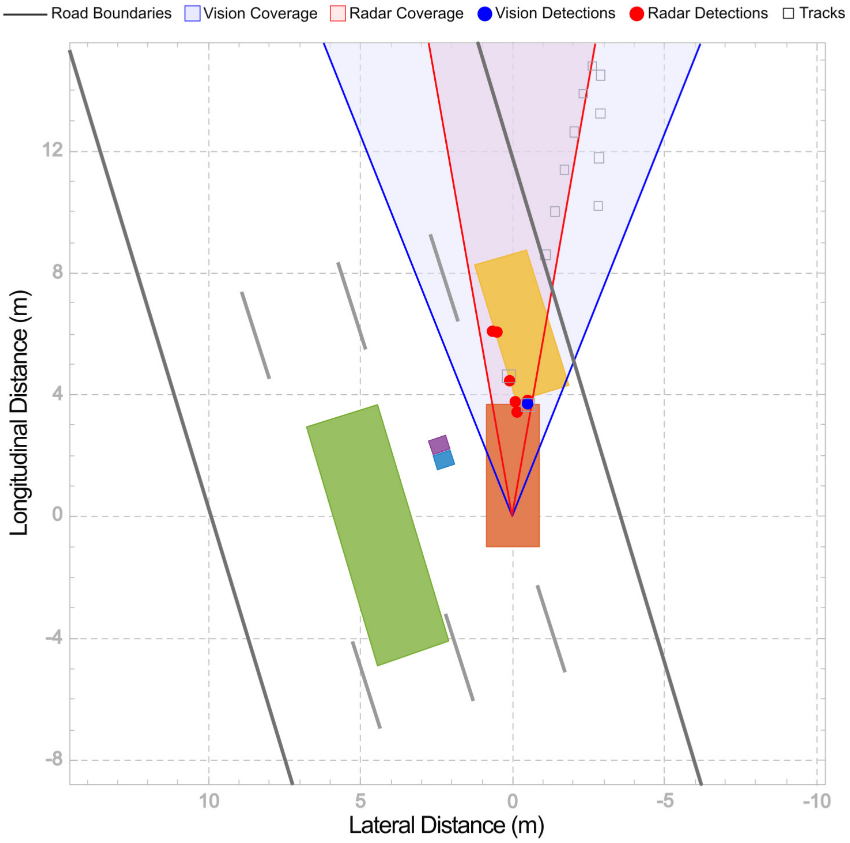

Figure 5.15: Distributive justice theory result scenario IIIA.The manoeuvre where less people is involved and has the smallest severity score is turning to the right.

### Agent with distributive justice beliefs

For the Distributive Justice theory in scenario IIIA, the agent calculates that the collision severity for the left side is higher than for the right side. In this situation, the number of people affected is one for the left and right turns, and three if the car continues without changing direction, so it decides to turn to the right, the direction with the smallest severity score and only one person affected, as shown in the agent console 5.9. This is illustrated by Fig. 5.15.

Agent console 5.9: Distributive justice agent, scenario IIIA

```
 --manner of collision--
 rtRE
 lftHO
 --Crash Severity--
rtCS: 0.027 lftCS0.072
**Distributive justice DM**
 People affected if turn to the left: 1
 People affected if turn to the right: 1
 People affected if not change direction: 3
 Severity score
 RtCS: 1.277, LFftCS: 1.322
[ethicalAgent] Risk of collision detected, steering to the right
```

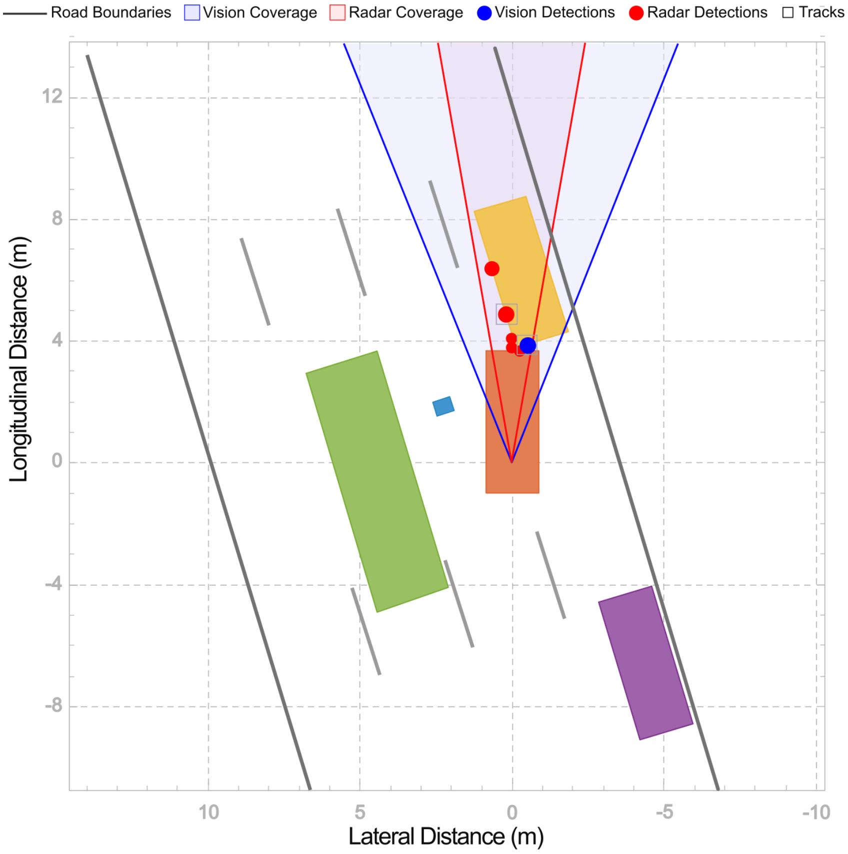

Figure 5.16: Distributive justice theory result scenario IIIB. In this scenario configuration, even though there are more people involved in the collision, if it turns right, this option is still the best one as it involves less people than going forward and it severity score is smaller that going left.

With scenario IIIB, the decision of steering to the right is reached again. In this situation the number of people affected has increased to two for the left and right sides, and remains at three if it does not change direction. Event though more people is affected in this situation, turning right is still the best option as shown by the severities scores calculated in the Agent console 5.9.

Agent console 5.10: Distributive justice agent, scenario IIIB

```
 --manner of collision--
 rtRE
 lftHO
**Distributive justice DM**
 People affected if turn to the left: 2
 People affected if turn to the right: 2
 People affected if not change direction: 3
 Severity score
 RtCS: 2.277, LFftCS: 2.322
[ethicalAgent] Risk of collision detected, steering to the right
```

### Agent with Kantian beliefs

In the Kantian mode there is no real difference for scenarios IIIA and IIIB. In both scenarios the agent first checks if there is a direction where it can avoid all the pedestrians. The agent knows that it needs to turn left or right to do it. The next step is to calculate the severity score for both directions, as seen in the agent console 5.11 for scenario IIIA and the Agent console 5.12 for scenario IIIB. From this it decides that the best option is to turn to the right. This is shown in Figures 5.17 and 5.18.

Agent console 5.11: Kantian agent, scenario IIIA

```
 --manner of collision--
 rtPV
 lftHO
**Kantian DM**
Can avoid pedestrians left
Can avoid pedestrian right
Can not avoid pedestrians front
 Severity score
RtCS: 0.012, LFftCS: 0.322
[ethicalAgent] Risk of collision detected, steering to the right
```

Agent console 5.12: Kantian agent, scenario IIIB

```
 --manner of collision--
 rtPV
 lftHO
 --Crash Severity--
rtCS: 0.027 lftCS0.072
**Kantian DM*
Can avoid pedestrians left
Can avoid pedestrian right
Can not avoid pedestrians front
 Severity score
RtCS: 0.012, LFftCS: 0.322
*[ethicalAgent] Risk of collision detected, steering to the left
```

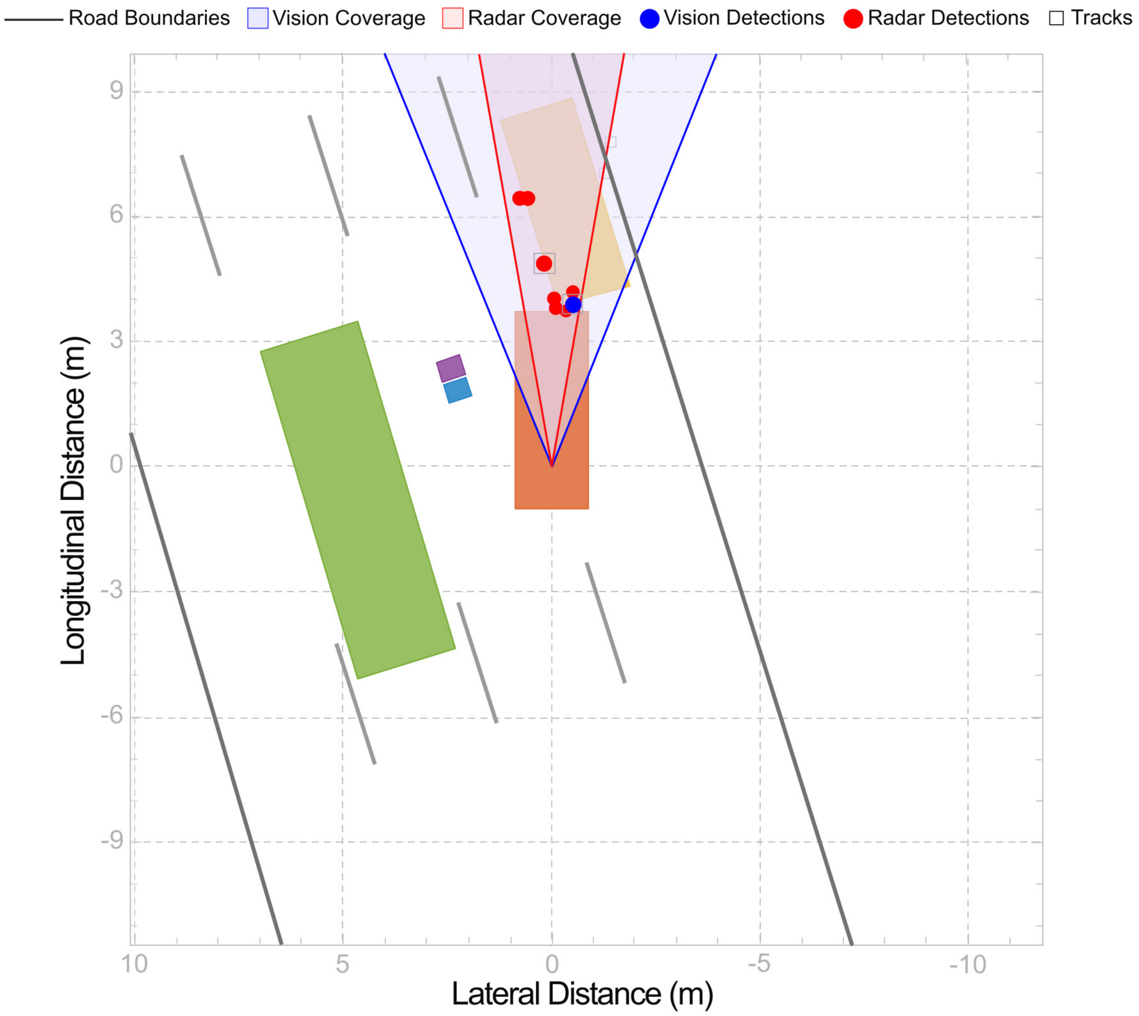

Figure 5.17: Kantian theory result scenario IIIA.

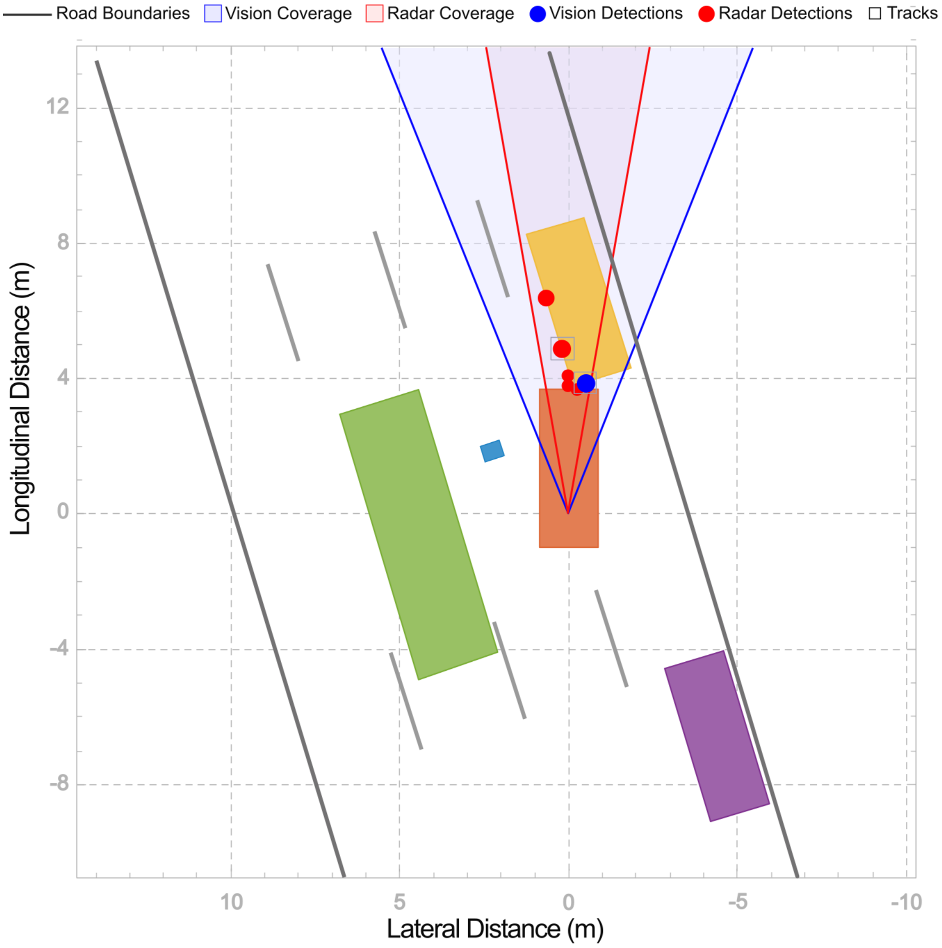

Figure 5.18: Kantian theory result scenario IIIB.

### Agent with altruist beliefs

In the Altruism mode, the priority is always the pedestrians. For that reason in scenarios IIIA and IIIB the agent reaches the same decision in both cases as shown in Fig. 5.19 and Fig. 5.20 this is because for this ethical theory it does not matter how many passengers are in the vehicle, it will always try to protect the POTV. For this it defines the direction in which is possible to avoid the pedestrians and then decides based on the severity score for each direction. In this case the better option is turn to the right as seen in the Agent console 5.13 for scenario IIIA and the Agent console 5.14 for scenario IIIB.

Agent console 5.13: Altruist agent, scenario IIIA

```
 --manner of collision--
 rtPV
lftHO
**altruist DM**
Can avoid pedestrians left
Can avoid pedestrian right
Can not avoid pedestrians front
 Severity score
 RtCS: 0.012, LFftCS: 0.25
[ethicalAgent] Risk of collision detected, steering to the right
```

Agent console 5.14: Altruist agent, scenario IIIB

```
 --manner of collision--
 rtPV
 lftHO
**altruist DM**
Can avoid pedestrians left
Can avoid pedestrian right
Can not avoid pedestrians front
  Severity score
 RtCS: 0.012, LFftCS: 0.25
[ethicalAgent] Risk of collision detected, steering to the right
```

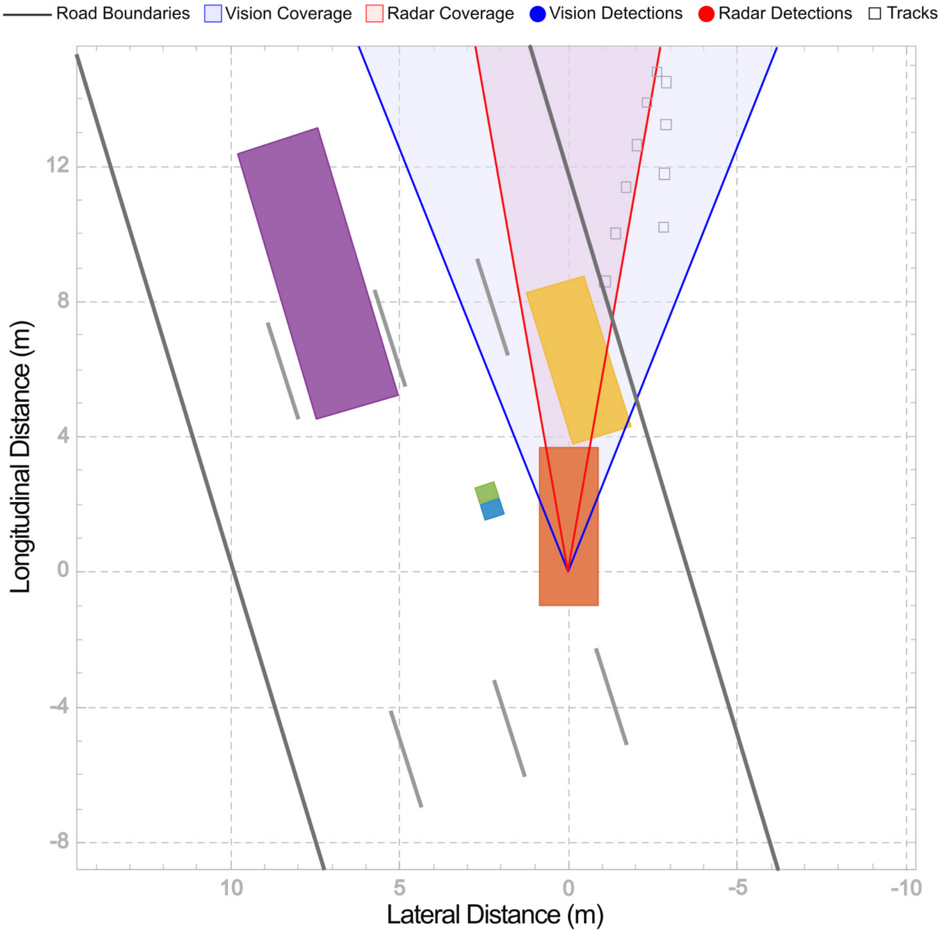

Figure 5.19: Altruist theory result scenario IA.

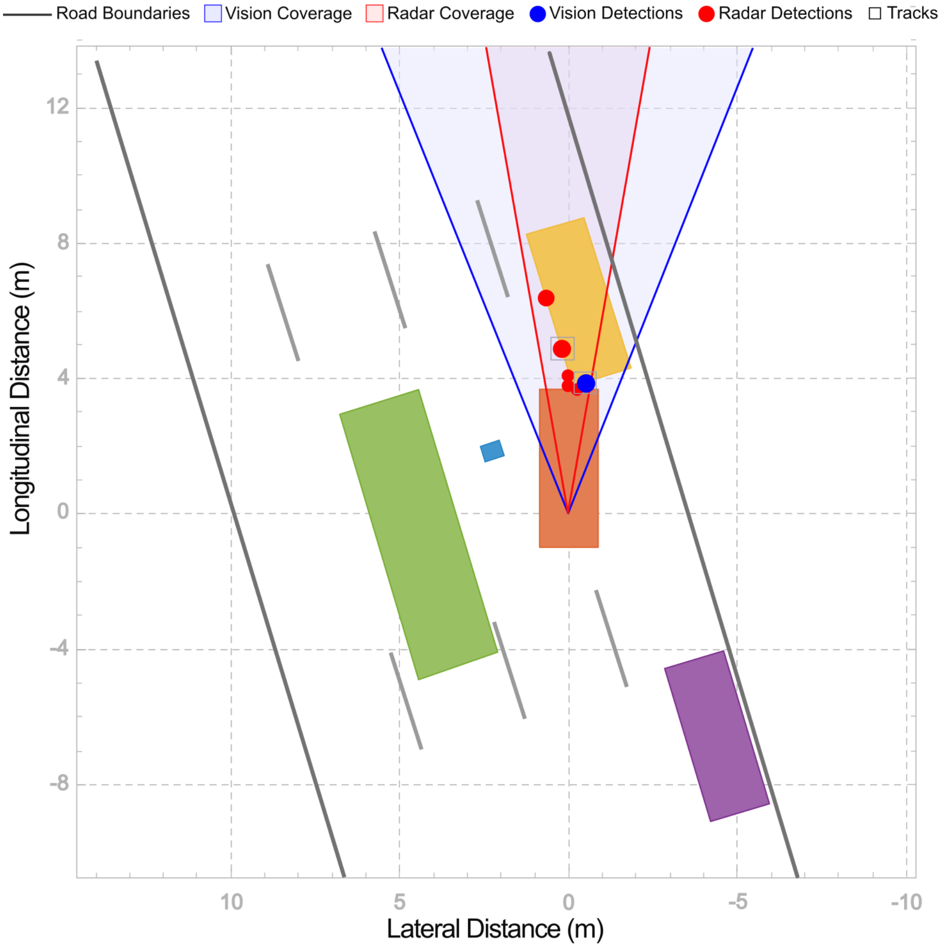

Figure 5.20: Altruist theory result scenario IB.

**Agent with ethical egoism beliefs**

The ethical egoism mode in scenario IIIA is not applicable as reflected in the agent console 5.15. This is because of the legal requirements implemented for the agent, in which it is not possible to harm two or more people for the sake of one. In this situation the agent prioritizes to avoid the pedestrians and calculates the severity score for each possible direction. In this case is safer to turn to the right as shown in Figure 5.21.

Agent console 5.15: Egoist agent, scenario IIIA

```
 --manner of collision--
 rtPV
 lftHO
**egoist DM**
There are more POTV than PAX egoist setting is not possible
Severity score
RtCS: 0.012, LFftCS: 0.25
[ethicalAgent] Risk of collision detected, steering to the right
```

In scenario IIIB there are more passengers on board of the vehicle than pedestrians, hence the egoistical mode can be used, the Agent console 5.16 shows how this agent behaves. In this instance it calculates the severity score for left and right and it decides not to change direction because it can not avoid the pedestrian without causing harm to the passengers, this result is shown in Figure 5.22.

A summary of the outcomes of all ethical settings in each scenario is shown in Table 5.5.

Agent console 5.16: Egoist agent, scenario IIIB

```
 --manner of collision--
 rtPV
 lftHO
**egoist DM**
Severity score
RtCS: 0.012, LFftCS: 0.25
[ethicalAgent] Risk of collision detected, keep driving
```

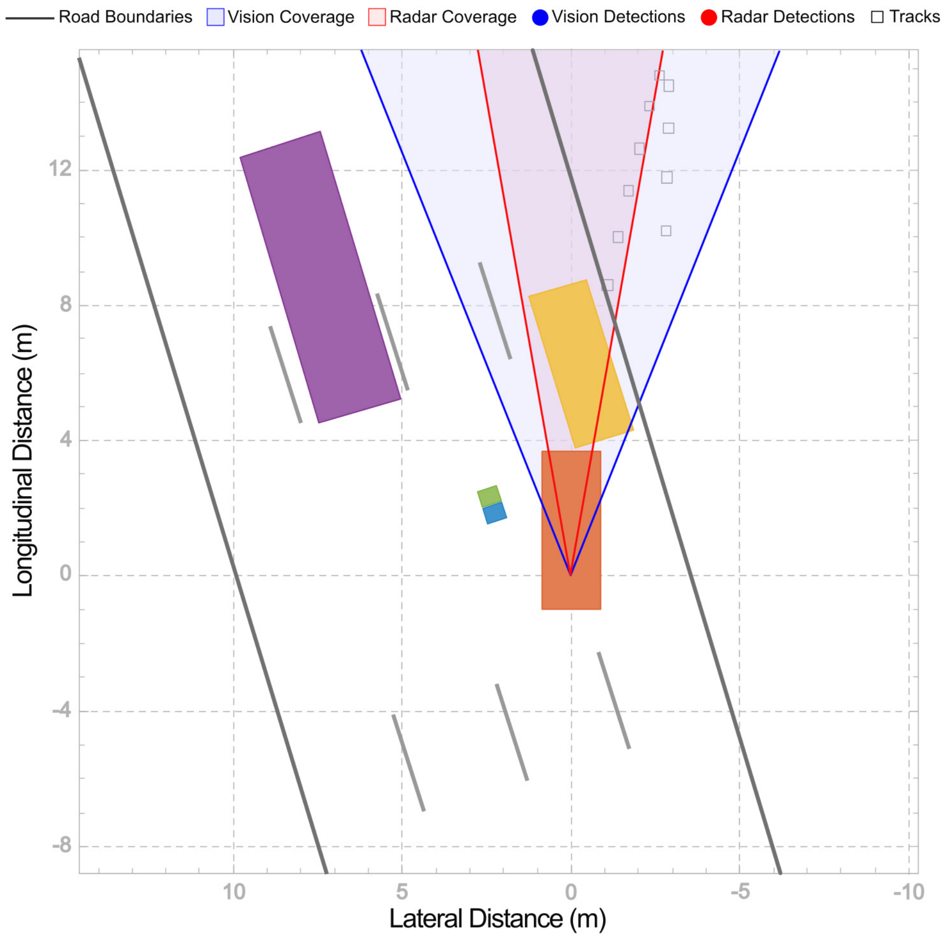

Figure 5.21: Ethical Egoism theory result scenario IIIA.

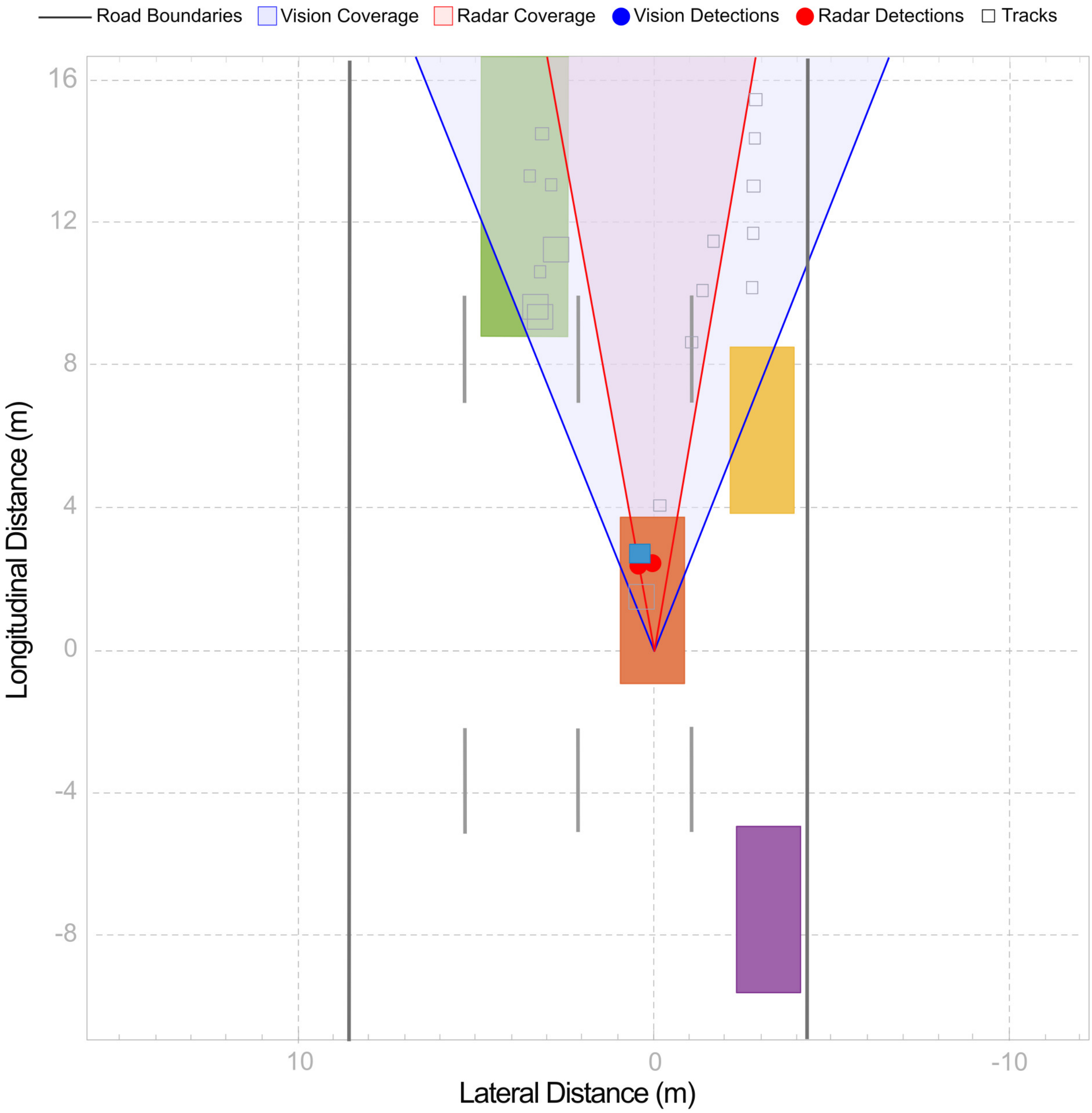

Figure 5.22: Ethical Egoism result scenario IIIB.

Table 5.5: Final results for all ethical agents in all the different scenarios.

| | Utilitarian | Distributive Justice | Kantian | Altruist | Egoist |
|---|---|---|---|---|---|
| Scenario I | Turns left<br>Avoids POTV | Turns left<br>Avoids POTV | Turns left<br>Avoids POTV | Turns left<br>Avoids POTV | Turns left<br>Avoids POTV |
| Scenario II | Keeps direction<br>Crashes with one POTV | Turns left<br>Crashes with other vehicle | Turns left<br>Crashes with other vehicle | Turns left<br>Crashes with other vehicle | Keeps direction<br>Crashes with one POTV |
| Scenario IIIA | Turns right<br>Crashes with parked vehicle | Turns right<br>Crashes with parked vehicle | Turns right<br>Crashes with parked vehicle | Turns right<br>Crashes with parked vehicle | Turns right<br>Crashes with parked vehicle |
| Scenario IIIB | Keeps direction<br>Crashes with one POTV | Turns right<br>Crashes with parked vehicle | Turns right<br>Crashes with parked vehicle | Turns right<br>Crashes with parked vehicle | Keeps direction<br>Crashes with one POTV |

### 5.4.4 Alternative Severity Score Calculation

To explore the agent behaviour when calculating the severity score via the multiplication of $\alpha$ and $\beta$, we carried out a simulation of the utilitarian mode in scenario II. The Agent console 5.17 shows the result. First, the agent determines that the priority is to protect the passengers, in line with the utilitarian approach as now there are more in number than the pedestrians. In this case, it is not possible to avoid the pedestrian without harming the passengers, and even with different values for the severity score, the agent decides to keep driving without changing direction in order to protect the passengers.

Agent console 5.17: Utilitarian agent, scenario IIIB with alternative severity score

```
 --manner of collision--
 rtPV
 lftHO
**utilitarian DM**
 Priority: protect PAX
 Severity score
 RtCS: 0.0036, LFftCS: 0.021
[ethicalAgent] Risk of collision detected, keep driving
```

After doing the same test in all the scenarios with the utilitarian mode, the results obtained are the same in both outcomes and decision-making processes, as in the original proposed method. However, the methodology involved in the calculation of the severity score is still an open question that requires further research, since a more detailed set of statistic data could improve the knowledge about possible outcomes and in consequence the way the score is calculated.

## 5.5 Discussion

As can be seen from the outcomes obtained, for some theories the end result is the same. This can be attributed to two factors. Firstly, to minimise the complexity of the study, the vehicle is restricted to three actions: Turn left, turn right and not change direction. The second factor is that because of the limited number of manoeuvres, the possibilities of arriving to the same conclusion are high. However, through the agent console we can analyse the decision-making process to assess if the agent is following the concerns defined in Table 3.1. Is worth nothing that even though the final manoeuvre can lead to the same outcome, the most important aspect of the ethical agent is the logical processing of the environment and the decision-making progress. This is important as people often question why these decisions were made and how these align with their ethical views. The outcomes of the scenarios and results presented here can help in the adoption of autonomous vehicles and can aid in the creation of new legislation, needed in this area.

Analysing the outcomes for the utilitarian approach, it is noted how its priority is to harm the least possible number of people, in line with the concerns defined in Table 3.1. This contrast is clear from the results in scenarios IIIA and IIIB. Scenario II shows how it avoids harming the passengers, as they are larger in number, and since there is not an option where it can avoid all the pedestrians, it keeps moving forward and harms the pedestrian that crossed its way without involving the bystander. Outcomes like this one can be a source of debate as it can be seen as unacceptable to harm the pedestrians. These are ethical considerations that need to be discussed with authorities, law makers, experts on ethics and autonomous vehicles and with the society in general. It is also important to note that some outcomes might be acceptable on certain countries and not in others, depending on the ethical views predominant in those places.

In scenarios IIIA and IIIB and II, the Kantian agent behaves as expected, always avoiding the pedestrian. In Scenario II, it is observed how the agent selects the option with the higher severity score because it is prioritizing its principal concern to avoid pedestrians as defined in Table 3.1.

The altruist agent outcomes for all scenarios shows how it always prioritizes protecting the pedestrians, no matter how many people are travelling in the vehicle, in line with its concerns as shown in Table 3.1. Once the agent has avoided all the pedestrians, proceeds to select the manoeuvre that carries the smallest severity score for the passengers.

For the ethical egoism theory, in scenario IIIA, the legal requirement of not harming two or more people to protect only one is shown. In the scenarios where this setting is applicable, it is observed that the agent always tries to minimise damage to the passengers, which is its main concern according to Table 3.1.

For the third scenario, all the outcomes obtained were the same. This show us that for all the ethical theories, the agent will always select the safest option. In this case, selecting the manoeuvre that in first instance can avoid the collision is the most preferred.

Although the results obtained from these tests are promising, the system would benefit from more exhaustive tests and adding more capability to the sensors in the vehicle. Currently, the ego vehicle operates with only one radar sensor and one camera, giving us only partial coverage of the vehicle surroundings. This minimal configuration was selected for simplicity, assuming that all the important events are happening in the front of the vehicle. Hence, a pair of front-facing radar and vision sensors are the minimum to operate the system and suitable for the presented scenarios. The testing scenarios were designed to account for this sensing limitation, however, more testing with superior sensing capabilities and more actors would prove valuable to further validate the ethical settings presented here. In regards to more scenarios, the agent could be tested in the ones used in the *Moral Machine* experiment presented in [76, 77]. In these scenarios, the autonomous vehicle must take decisions considering physical and personal attributes about the people involved. It would be interesting to see how our agent behaves.

Due to time constraints, these simulations were limited to a single vehicle that operates with the ethical agent. This is fine considering that although every day there may be more autonomous vehicles in operation, common vehicles are not going to be replaced immediately, generating situations such as those presented in these tests, in which the autonomous vehicle is the only one with these capabilities in certain situations. However, thinking ahead, testing scenarios with more than one ethical agent can offer us many more possibilities in terms of the development of autonomous vehicles. Having more than one agent in the scenario could allow the ego vehicle to make assumptions about how the other vehicle could react based on the different ethical settings. Another possibility worth exploring would be testing with vehicles with communication capabilities that allow them to evaluate the scenario and make a joint decision.

## 5.6 Conclusion

In this chapter we presented the scenarios designed to test the system proposed in Chapter 4. For this purpose, three scenarios were created and each of the ethical settings was run. From the results obtained, we can conclude that the agent is behaving in an expected way, according to the concerns expressed in Section 3.2, Table 3.1. From the outcomes in the proposed scenarios, we can conclude that for each ethical setting, the agent outputs a decision-making process that aligns with the corresponding ethical concerns. In other words, we can say that the agent captures each ethical setting successfully when presented with a dilemma that requires a decision.

The results presented here are promising in the sense that there were successful in fulfilling the expected behaviour of the agent according to Table 5.5. However, more tests are required to fully guarantee the correct operation of the system before a real-life implementation.

# Chapter 6

# Conclusions

This thesis presents research in ethical decision-making for inevitable collisions using BDI agents. In this chapter, we present a discussion about this work, results, general conclusions and future directions.

## 6.1 Conclusions and discussion

The implementation of ethics in decision-making for AVs is a subject that has gained more importance in the past few years as AVs are becoming more of a reality. The principal aim of this work was to develop a system that could provide different ethical approaches for autonomous vehicles decision-making processes when they encounter an unavoidable collision. In Chapter 3 we studied different ethical theories that could be applied to AVs. We also defined eight ethical concerns for the operation of the vehicle. These concerns were ranked in order of importance based on each ethical theory definition. From Table 3.1 it is clear that concerns differ in importance according to the different theories. Interestingly, no matter what ethical theory we select there are concerns that always are more important than others, i.e. protecting human lives is always more important than animals or objects. An advantage of a system like this is that it can provide more than one solution to cater for different preferences and without forcing one single solution on everyone within a broader legislative framework.

In Chapter 4 we presented a BDI agent to implement the ethical concerns defined in Chapter 3. The agent implemented has a hybrid architecture, with an abstraction layer between the agent and the real world. The agent receives information about the world as perceptions and then uses this information to estimate the *Crash Severity*. This estimate is a calculation which takes into account the static location of the actors, the manner of collision and the type of objects involved in the collision. The agent then uses this calculation to make a decision. The decision is informed to the vehicle control in the form of a command to engage braking or a direction for steering.

Chapter 5 covers the implementation of the vehicle control and the simulation of the environment. Three scenarios were created to test the ethical agent behaviour. From the results obtained, we can conclude that each ethical setting is handling the different scenarios in line with the ethical concerns defined in Chapter 3, with each ethical setting prioritizing its corresponding most important concern. As shown in Table 5.5, the outcomes for the same scenario in each ethical setting is sometimes the same. This is expected as the manoeuvres available to the vehicle are limited, however, the difference lies in the way in which the agent selected that action. Knowing the decision process could help us to understand and accept why the decision was taken and is one of the main contributions of this work.

## 6.2 Societal impact

According to the World Health Organization, road accidents are the 8th cause of death worldwide and the leading cause of death for people in the group aged 6 to 29 years [78]. The inclusion of autonomous vehicles could help to improve this situation by reducing the number of road accidents due to human error. Other benefits for society include an improvement in transportation options for people with reduced mobility or with any impairment that prevents them from driving a regular vehicle. Also, the improvement of traffic management and increased efficiency on fuel consumption will be beneficial for the environment [79]. However, in order for AVs to be widely available, there is the need for public acceptance and willingness to buy this type of vehicle.

The major proposed contribution of this research is to provide the industry and the public with a different option on how to tackle the ethical questioning surrounding autonomous vehicles. Research shows that people from different cultures have different ethical positions [80, 81]. Out proposal enables the user, i.e. the passengers of the vehicle, to select an ethical vision that aligns with their beliefs, while at the same time complying with relevant laws. This could improve the reception of AVs. As noted in [25, 26], humans show aversion to machines taking an ethical decision and do not completely agree with a legally enforced utilitarian approach [4]. Involving the user of the autonomous vehicle in the decision making process according to its ethical view could result in a wider acceptability of AVs.

## 6.3 Limitations and Future Directions

In this work we have assumed that there is no uncertainty in our system, we do not take into account any sensor, environmental or dynamics uncertainty. This is an important area in which this work could be further expanded. There is research suggesting how uncertainties can be accounted for in autonomous systems in different ways like probabilistic approaches [82, 83] or MPC control, where the uncertainty is introduced in the modelling phase, then the model is calibrated to account for uncertainties with data collected during driving [84]. In the literature, we can also find approaches to handle uncertainties from an ethical perspective, [85] presents a framework to do this from a Kantian perspective, they propose a set of rules that act as a threshold in which risk to harm a person can be ignored. Although it might only be applicable in certain situations, due to the nature of Kantian ethics. Some argue that if a decision violates the principles of Kantian ethics this decision should not be made. However, this is not always practical. Hence, in addition to handling uncertainties according to each ethical theory, another approach that can be explored is to analyse how the existence of uncertainty impacts the acceptance of ethical decision-making by autonomous vehicles. In [86], the authors investigate what AV behaviours people find morally acceptable when risk and uncertainty are present. According to their findings, when presented with an scenario where the options are staying in their lane and collide with a pedestrian, or swerve and collide with a bystander, people's preference was to stay on the lane, even in situations where the probability of harming the pedestrian was high and the probability of colliding with the bystander was low or uncertain. In conclusion, people consider that in uncertain situations it is more acceptable not to take action and let the situation follow its course than take action and cause an accident.

Another unexplored aspect in this research is how the ethical preferences would be elicited from the user. We assume that the vehicle possesses a graphical interface, i.e. an onboard screen, where the ethical preferences could be inputted. However, a study solely focused on this would be necessary since the graphical interface should be able to clearly communicate the ethical theories available and the implications that each of them could have. Further research can focus on the development of a user interface to explore people perception and level of acceptance of this kind of implementation. This could be done through studies with people from different backgrounds, presenting them with different scenarios, and asking them to select an ethical theory, measuring the level of understanding of the interface and their acceptance of the outcome.

Due to time limitations, we could not expand this work to be able to have two ethical agents implemented together. Doing this in the future would give an opportunity to better understand how the ethical agent would behave under the assumption that there are more agents with the same decision-making capabilities. Additionally, the aspect of multi agent communication could also be explored, as it might help to improve the safety of all people involved in an accident scenario.

Regarding the legal part of the system, in the future, the system could be improved by tailoring it to the legal requirements in different countries, as these are not always the same, and they can vary even within the same country [87]. Testing the system with different legal requirements could help us understand if and how this can impact the ethical decision-making process.

Another aspect that we can mention is that the ethical concerns presented in Chapter 3 and published in [88] were developed from an engineering point of view. Although our ethical interpretations have been discussed with some ethicists, these could benefit from a broader discussion within the philosophy field to assess the soundness of our approach.

Lastly, the work presented in this thesis is limited to computational simulations. As future work, once the limitations indicated in this section have been addressed, the next step could be to test the system in a physical vehicle within a safe and controlled environment.

# Bibliography

[1] R. H. Bordini, J. F. Hübner, and M. J. Wooldridge, *Programming multi-agent systems in AgentSpeak using Jason*. J. Wiley, 2007.

[2] N. Goodall, "Ethical Decision Making During Automated Vehicle Crashes," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2424, no. 2424, pp. 58–65, 2014.

[3] J. Guo, U. Kurup, and M. Shah, "Is it Safe to Drive? An Overview of Factors, Metrics, and Datasets for Driveability Assessment in Autonomous Driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3135–3151, aug 2019.

[4] J.-F. Bonnefon, A. Shariff, and I. Rahwan, "The social dilemma of autonomous vehicles," *Science*, vol. 352, no. 6293, pp. 1573–1576, jun 2016.

[5] L. T. Bergmann, L. Schlicht, C. Meixner, P. König, G. Pipa, S. Boshammer, and A. Stephan, "Autonomous Vehicles Require Socio-Political Acceptance—An Empirical and Philosophical Perspective on the Problem of Moral Decision Making," *Frontiers in Behavioral Neuroscience*, vol. 12, no. February, pp. 1–12, 2018.

[6] R. Noothigattu, S. N. S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, and A. D. Procaccia, "A Voting-Based System for Ethical Decision Making," sep 2017.

[7] P. Lin, "Why Ethics Matters for Autonomous Cars," in *Autonomous Driving: Technical, Legal and Social Aspects*, M. Maurer, J. C. Gerdes, B. Lenz, and H. Winner, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, ch. 4, pp. 69–85.

[8] U. Bose, "An ethical framework in information systems decision making using normative theories of business ethics," *Ethics and Information Technology*, vol. 14, no. 1, pp. 17–26, 2012.

[9] C. Fisher and A. Lovell, *Business ethics and values : individual, corporate and international perspectives*, 2nd ed. Pearson Education Limited, 2006.

[10] C. E. Johnson, *Organizational ethics: A practical approach*, 2nd ed. United States of America: SAGE Publications, 2012.

[11] J. Rawls, *A theory of justice*. Cambridge Mass.: Belknap Press of Harvard University Press, 1971.

[12] M. Boylan, *Basic Ethics*, 2nd ed. Pearson/Prentice Hall, 2009.

[13] C. Yoon, "Ethical decision-making in the Internet context: Development and test of an initial model based on moral philosophy," *Computers in Human Behavior*, vol. 27, no. 6, pp. 2401–2409, 2011.

[14] P. Foot, "The Problem of Abortion and the Doctrine of Double Effect," *Oxford Review*, vol. 5, pp. 5–15, 1967.

[15] N. J. Goodall, "Can you program ethics into a self-driving car?" *IEEE Spectrum*, vol. 53, no. 6, pp. 28–58, jun 2016.

[16] E. commission, "Automated and connected driving report," Federal Minister of Transport and Digital Infrastructure, Tech. Rep., 2017. [Online]. Available: https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission-automated-and-connected-driving.pdf?{_}{_}blob=publicationFile

[17] IEEE, "IEEE Code of Ethics." [Online]. Available: https://www.ieee.org/about/corporate/governance/p7-8.html

[18] T. Fournier, "Will My Next Car Be a Libertarian or a Utilitarian?: Who Will Decide?" *IEEE Technology and Society Magazine*, vol. 35, no. 2, pp. 40–45, jun 2016.

[19] J. Gogoll and J. F. Müller, "Autonomous Cars: In Favor of a Mandatory Ethics Setting," *Science and Engineering Ethics*, vol. 23, no. 3, pp. 681–700, 2017.

[20] H. Y. Liu, "Irresponsibilities , inequalities and injustice for autonomous vehicles," *Ethics and Information Technology*, vol. 19, no. 3, pp. 193–207, 2017.

[21] S. Nyholm and J. Smids, “The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?” *Ethical Theory and Moral Practice*, vol. 19, no. 5, pp. 1275–1289, 2016.

[22] R. Benenson, S. Petti, T. Fraichard, and M. Parent, “Special Issue on Advances in Autonomous Vehicle Technologies for Urban Environment,” *International Journal of Vehicle Autonomous Systems, Inderscience*, vol. 1, no. 2, pp. 4–23, 2008. [Online]. Available: https://hal.inria.fr/inria-00115112v2

[23] J. Li, X. Zhao, M.-J. Cho, W. Ju, and B. F. Malle, “From Trolley to Autonomous Vehicle: Perceptions of Responsibility and Moral Norms in Traffic Accidents with Self-Driving Cars,” 2016.

[24] Y. E. Bigman and K. Gray, “People are averse to machines making moral decisions,” *Cognition*, vol. 181, no. August, pp. 21–34, 2018.

[25] J. Gogoll and M. Uhl, “Rage against the machine: Automation in the moral domain,” *Journal of Behavioral and Experimental Economics*, vol. 74, no. March, pp. 97–103, 2018.

[26] M. König and L. Neumayr, “Users’ resistance towards radical innovations: The case of the self-driving car,” *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 44, pp. 42–52, jan 2017.

[27] N. Gold, A. M. Colman, and B. D. Pulford, “Cultural differences in responses to real-life and hypothetical trolley problems.” *Judgment and Decision Making*, vol. 9, no. 1, pp. 65–76, 2014.

[28] D. Leben, “A Rawlsian algorithm for autonomous vehicles,” *Ethics and Information Technology*, vol. 19, no. 2, pp. 107–115, jun 2017.

[29] G. Contissa, F. Lagioia, and G. Sartor, “The Ethical Knob: ethically-customisable automated vehicles and the law,” *Artificial Intelligence and Law*, vol. 25, no. 3, pp. 365–378, sep 2017.

[30] P. Lin, “The Ethics of Saving Lives With Autonomous Cars Is Far Murkier Than You Think,” 2013. [Online]. Available: https://www.wired.com/2013/07/the-surprising-ethics-of-robot-cars/

[31] A. Hevelke and J. Nida-Rümelin, “Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis,” *Science and Engineering Ethics*, vol. 21, no. 3, pp. 619–630, 2015.

[32] S. Maki and A. Sage, “Self-driving Uber car kills Arizona woman crossing street - Reuters,” 2018. [Online]. Available: https://www.reuters.com/article/us-autos-selfdriving-uber/self-driving-uber-car-kills-arizona-woman-crossing-street-idUSKBN1GV296

[33] Shepardson David, “Third fatal Tesla Autopilot crash renews questions about system - Reuters,” 2019. [Online]. Available: https://www.reuters.com/article/us-tesla-autopilot/third-fatal-tesla-autopilot-crash-renews-questions-about-system-idUSKCN1SM1QE

[34] T. Chen, F. Chen, and C. Chen, “Review on Driverless Traffic from Management Perspective,” *MATEC Web of Conferences*, vol. 124, p. 03002, 2017.

[35] V. Ilkova and A. Ilka, “Legal aspects of autonomous vehicles-An overview,” *Proceedings of the 2017 21st International Conference on Process Control, PC 2017*, pp. 428–433, 2017.

[36] S. Gless, E. Silverman, and T. Weigend, “If Robots cause harm, Who is to blame? Self-driving Cars and Criminal Liability,” *New Criminal Law Review: An International and Interdisciplinary Journal*, vol. 19, no. 3, pp. 412–436, 2016.

[37] A. K. Taylor and S. Bouazzaoui, “Moving Forward with Autonomous Systems: Ethical Dilemmas,” in *Advances in Human Factors and Systems Interaction*, ser. Advances in Intelligent Systems and Computing, I. L. Nunes, Ed. Cham: Springer International Publishing, 2019, vol. 781, pp. 101–108.

[38] F. Santoni de Sio, “Killing by Autonomous Vehicles and the Legal Doctrine of Necessity,” *Ethical Theory and Moral Practice*, vol. 20, no. 2, pp. 411–429, 2017.

[39] I. Coca-Vila, “Self-driving Cars in Dilemmatic Situations: An Approach Based on the Theory of Justification in Criminal Law,” *Criminal Law and Philosophy*, vol. 12, no. 1, pp. 59–82, mar 2018.

[40] A. M. Khan, A. Bacchus, and S. Erwin, “Policy challenges of increasing automation in driving,” *IATSS Research*, vol. 35, no. 2, pp. 79–89, 2012.

[41] J. Lederman, M. Garrett, and B. D. Taylor, “Fault-y Reasoning: Navigating the Liability Terrain in Intelligent Transportation Systems,” *Public Works Management and Policy*, vol. 21, no. 1, pp. 5–27, 2016.

[42] J. Fleetwood, “Public Health, Ethics, and Autonomous Vehicles,” *American Journal of Public Health*, vol. 107, no. 4, pp. 532–537, apr 2017.

[43] R. Schubert, K. Schulze, and G. Wanielik, “Situation Assessment for Automatic Lane-Change Maneuvers,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 3, pp. 607–616, sep 2010.

[44] R. Schubert, “Evaluating the utility of driving: Toward automated decision making under uncertainty,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 1, pp. 354–364, 2012.

[45] Y. Hou, P. Edara, and C. Sun, “Situation assessment and decision making for lane change assistance using ensemble learning methods,” *Expert Systems with Applications*, vol. 42, no. 8, pp. 3875–3882, 2015.

[46] M. Ardelt, P. Waldmann, F. Homm, and N. Kaempchen, “Strategic decision-making process in advanced driver assistance systems,” *IFAC Proceedings Volumes (IFAC-PapersOnline)*, no. Grünheid 2008, pp. 566–571, 2010.

[47] J. Nilsson, J. Silvlin, M. Brannstrom, E. Coelingh, and J. Fredriksson, “If, When, and How to Perform Lane Change Maneuvers on Highways,” *IEEE Intelligent Transportation Systems Magazine*, vol. 8, no. 4, pp. 68–78, 2016.

[48] L. Zhao, R. Ichise, T. Yoshikawa, T. Naito, T. Kakinami, and Y. Sasaki, “Ontology-based decision making on uncontrolled intersections and narrow roads,” *IEEE Intelligent Vehicles Symposium, Proceedings*, vol. 2015-Augus, no. Iv, pp. 83–88, 2015.

[49] M. Pellkofer and E. Dickmanns, “Behavior decision in autonomous vehicles,” in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 2. IEEE, 2002, pp. 495–500.

[50] S. D. Kim, C. W. Roh, S. C. Kang, and J. B. Song, "A fuzzy decision making algorithm for safe driving in urban environment," *ICCAS 2007 - International Conference on Control, Automation and Systems*, pp. 678–683, 2007.

[51] X.-m. Chen, G. Tian, C.-y. Chan, Y.-s. Miao, J.-w. Gong, and Y. Jiang, "Bionic Lane Driving Decision-Making Analysis for Autonomous Vehicle Under Complex Urban Environment," *Transportation Research Board*, pp. 1–20, 2016.

[52] A. Furda and L. Vlacic, *Multiple criteria-based real-time decision making by autonomous city vehicles*. IFAC, 2010, vol. 7, no. PART 1.

[53] ——, "Enabling Safe Autonomous Driving in Real-World City Traffic Using Multiple Criteria Decision Making," *IEEE Intelligent Transportation Systems Magazine*, vol. 3, no. 1, pp. 4–17, 2011.

[54] A. Best, S. Narang, D. Barber, and D. Manocha, "AutonoVi: Autonomous vehicle planning with dynamic maneuvers and traffic constraints," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2017-Septe, pp. 2629–2636, 2017.

[55] M. A. Goodrich and E. R. Boer, "Designing Human-Centered Automation: Tradeoffs in Collision Avoidance System Design," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 1, pp. 40–54, 2000.

[56] J. Hillenbrand, A. Spieker, and K. Kroschel, "A Multilevel Collision Mitigation Approach mdash;Its Situation Assessment, Decision Making, and Performance Tradeoffs," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 4, pp. 528–540, 2006.

[57] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster, "Formal verification of ethical choices in autonomous systems," *Robotics and Autonomous Systems*, vol. 77, pp. 1–14, mar 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0921889015003000

[58] N. De Moura, R. Chatila, K. Evans, S. Chauvier, and E. Dogan, "Ethical decision making for autonomous vehicles," *IEEE Intelligent Vehicles Symposium, Proceedings*, pp. 2006–2013, 2020.

[59] K. Evans, N. de Moura, S. Chauvier, R. Chatila, and E. Dogan, "Ethical Decision Making in Autonomous Vehicles: The AV Ethics Project," *Science and Engineering Ethics*, vol. 26, no. 6, pp. 3285–3312, dec 2020. [Online]. Available: https://link.springer.com/article/10.1007/s11948-020-00272-8

[60] European Commission, "On the road to automated mobility: An EU strategy for mobility of the future," European Commission, Brussels, Tech. Rep., 2018. [Online]. Available: https://ec.europa.eu/transport/sites/transport/files/3rd-mobility-pack/com20180283{_}en.pdf

[61] National Highway Traffic Safety Administration, "Automated driving systems 2.0: a vision for safety," 2017. [Online]. Available: https://www.nhtsa.gov/manufacturers/automated-driving-systems

[62] E. L. Chao, "Automated driving systems 2.0: A vision for safety," NHTSA, Tech. Rep., 2017.

[63] L. A. Dennis, M. Fisher, N. K. Lincoln, A. Lisitsa, and S. M. Veres, "Declarative Abstractions for Agent Based Hybrid Control Systems," *Lecture Notes in Computer Science*, vol. 6619 LNAI, pp. 96–111, 2010. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-20715-0_6

[64] ——, "Practical verification of decision-making in agent-based autonomous systems," *Automated Software Engineering*, vol. 23, no. 3, pp. 305–359, sep 2016.

[65] M. Bratman, *Intention, Plans, and Practical Reason.* Cambridge: Cambridge, MA: Harvard University Press, 1987.

[66] G. Weiss, *Intelligent Agents*, 2000, pp. 27–77.

[67] D. N. Kinny and M. P. Georgeff, "Commitment and Effectiveness of Situated Agents," *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI- 91)*, pp. 82–88, 1991.

[68] A. S. Rao, "AgentSpeak(L): BDI agents speak out in a logical computable language," *Lecture Notes in Computer Science*, vol. 1038, pp. 42–55, 1996.

[69] L. Padgham and M. Winikoff, *Developing Intelligent Agent Systems A practical guide*. Chichester: Wiley, 2004.

[70] National Highway Traffic Safety Administration, "Fatality Analysis Reporting System (FARS) Encyclopedia: Help - Terms." [Online]. Available: https://www-fars.nhtsa.dot.gov/help/terms.aspx

[71] ——, "Traffic Safety Facts Annual Report Tables." [Online]. Available: https://cdan.nhtsa.gov/tsftables/tsfar.htm

[72] T. D. Gillespie, *Fundamentals of vehicle dynamics*. Society of Automotive Engineers, 1992.

[73] R. Rajamani, *Vehicle dynamics and control*. Springer, 2011.

[74] N. K. Lincoln, S. M. Veres, L. A. Dennis, M. Fisher, and A. Lisitsa, "Autonomous asteroid exploration by rational agents," *IEEE Computational Intelligence Magazine*, vol. 8, no. 4, pp. 25–38, 2013.

[75] "Audi A4 Saloon — A4 Range — Audi UK." [Online]. Available: https://www.audi.co.uk/uk/web/en/models/a4/a4-saloon.html#

[76] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J. F. Bonnefon, and I. Rahwan, "The Moral Machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, nov 2018.

[77] E. Awadm, S. Dsouza, I. Rahwan, A. Shariff, and J.-F. Bonnefon, "Moral Machine," p. http://moralmachine.mit.edu/, 2017. [Online]. Available: https://www.moralmachine.net/

[78] World Health Organization, "Road traffic injuries." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[79] National Highway Traffic Safety Administration, "The Evolution of Automated Safety Technologies." [Online]. Available: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety

[80] D. R. Forsyth, E. H. O'Boyle, and M. A. McDaniel, "East meets West: A meta-analytic investigation of cultural variations in idealism and relativism," *Journal of Business Ethics*, vol. 83, no. 4, pp. 813–833, dec 2008.

[81] J. Rhim, G. bbeum Lee, and J. H. Lee, "Human moral reasoning types in autonomous vehicle moral dilemma: A cross-cultural comparison of Korea and Canada," *Computers in Human Behavior*, vol. 102, pp. 39–56, jan 2020.

[82] S. Thrun, "Probabilistic algorithms in robotics," *The AI magazine*, vol. 21, no. 4, pp. 93–109, 2000.

[83] C. Hubmann, J. Schulz, M. Becker, D. Althoff, and C. Stiller, "Automated Driving in Uncertain Environments: Planning with Interaction and Uncertain Maneuver Prediction," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 1, pp. 5–17, mar 2018.

[84] A. Carvalho, S. Lefévre, G. Schildbach, J. Kong, and F. Borrelli, "Automated driving: The role of forecasts and uncertainty - A control perspective," vol. 24, pp. 14–32, jul 2015.

[85] A. Bjorndahl, A. J. London, and K. J. Zollman, "Kantian decision making under uncertainty: Dignity, price, and consistency," *Philosophers' imprint*, vol. 17, no. 7, pp. 1–22, apr 2017.

[86] B. Meder, N. Fleischhut, N. C. Krumnau, and M. R. Waldmann, "How Should Autonomous Cars Drive? A Preference for Defaults in Moral Judgments Under Risk and Uncertainty," *Risk Analysis*, vol. 39, no. 2, pp. 295–314, feb 2019.

[87] National Conference of State Legislatures, "Autonomous Vehicles — Self-Driving Vehicles Enacted Legislation." [Online]. Available: https://www.ncsl.org/research/transportation/autonomous-vehicles-self-driving-vehicles-enacted-legislation.aspx

[88] L. Millan-Blanquel, S. M. Veres, and R. C. Purshouse, "Ethical considerations for a decision making system for autonomous vehicles during an inevitable collision," *2020 28th Mediterranean Conference on Control and Automation, MED 2020*, pp. 514–519, sep 2020.